

# The Essence of Linear Algebra

## Introduction

### Most important concepts of linear algebra

<https://www.youtube.com/watch?v=YrHlHbtISM0&list=PLUl4u3cNGP61iQEFiWLE21EJCxwmWvvek&index=1>

$A = CR = \begin{bmatrix} & \\ & \end{bmatrix} \begin{bmatrix} & \\ & \end{bmatrix}$ Independent columns in $C$	
$A = LU = \begin{bmatrix} 0 \\ \diagdown \end{bmatrix} \begin{bmatrix} & \\ 0 \end{bmatrix}$ Triangular matrices $L$ and $U$	
$A = QR = \begin{bmatrix} q_1 & q_n \end{bmatrix} \begin{bmatrix} & \\ 0 \end{bmatrix}$ Orthogonal columns in $Q$	
$S = Q\Lambda Q^T \quad Q^T = Q^{-1}$ Orthogonal eigenvectors $Sq = \lambda q$	
$A = X\Lambda X^{-1}$ Eigenvalues in $\Lambda$ Eigenvectors in $X$ $Ax = \lambda x$	
$A = U\Sigma V^T$ Diagonal $\Sigma$ = Singular values $\sigma = \sqrt{\lambda(A^T A)}$ Orthogonal vectors in $U^T U = V^T V = I$ $Av = \sigma u$	

### Linearity means Homogeneity and Additivity

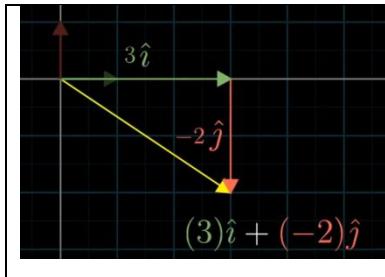
A system is linear if it is homogeneous and additive

<p><b>Homo</b> <b>gen</b></p> $f(\alpha x) = \alpha \cdot f(x)$ $f(\beta \cdot x) = \beta \cdot f(x)$ $f((\alpha + \beta)x) = (\alpha + \beta) \cdot f(x)$ $= \alpha \cdot f(x) + \beta \cdot f(x)$ <p><b>Additi</b> <b>vity</b></p> $f(\alpha x) = \alpha \cdot f(x)$ $f(\beta \cdot x) = \beta \cdot f(x)$ $f((\alpha + \beta)x) = (\alpha + \beta) \cdot f(x)$ $= \alpha \cdot f(x) + \beta \cdot f(x)$	<p>Suppose that you don't know the function <math>f</math> and you try to extract its behaviour by inputting values to it and measuring the output.</p> <p>A function with the graph of a line, is both homogeneous and additive. Every function that exhibits these two properties can be considered a linear function. These functions are of degree one in terms of their input <math>x</math> (<math>x^1</math>).</p>
<p><b>Homogeneity</b></p> <p><b>Additivity</b></p>	<p>In this case both the input and the output are functions of time <math>x(t)</math> and <math>y(t)</math> but the concept is the same. You can think of the input as if you input just a number (the value of the input at a specific time) and you examine the output number and repeat. From the relationship you can extract the behaviour of the function.</p>

### Scaling and adding

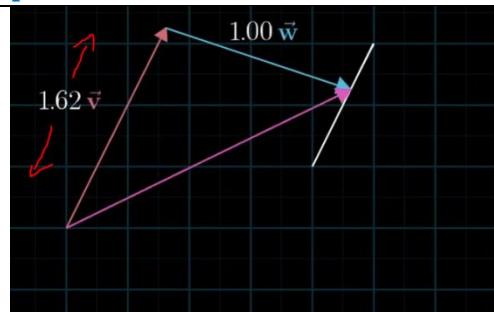
It doesn't matter if you think of vectors as arrows in space that happen to have a certain numerical representation or if you think of them as a list of numbers that happen to have a nice geometric interpretation. The usefulness of linear algebra has less to do with one of those definitions than it has with the ability to translate back and forth between them. (We will see in detail, that a vector can be anything where there is a sensible notion of **adding** two of them together and **multiplying** them by a number)

Multiplication of a vector by a number scales the vector, thus the numbers in linear algebra can be thought of as scalars (things that cause vectors to scale).

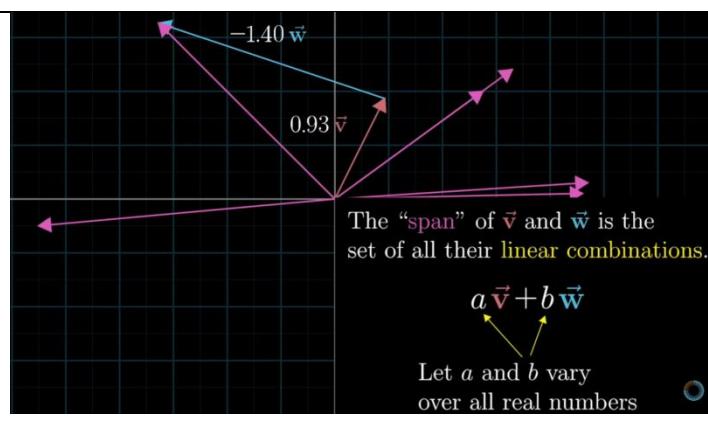


You can think of a vector's coordinates as scalars of the basis vectors. In this sense each vector is an addition of scaled basis vectors.

## Span and basis

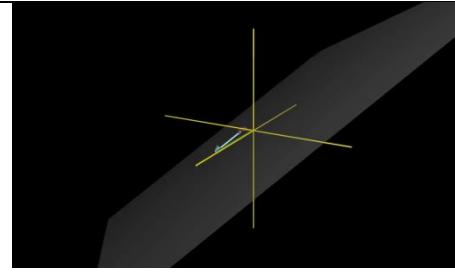


Scaling two vectors and adding them together is called a linear combination of these vectors. What does this have to do with lines? One way to think of it is that if you keep one scalar fixed and let the other change freely, the resulting vector tip will lie in a line



The span of most pairs of 2d vectors is ALL vectors of 2d space. But if the two vectors are aligned then their span is all vectors whose tip sits in this line.

The span of two vectors is a way of asking what are all the possible vectors you can reach using only these two fundamental operations, vector addition and scalar multiplication.



The span of two 3dimensional vectors is a 2d plane crossing the origin of the coordinate system.

The span of 3 3d vectors is the whole 3d space except from the case where two of the three are aligned or if one lies in the span of the other two, in which case the span is a 2d plane. In this case you can remove the third vector without affecting the span. In this case we say that the two vectors (from which you removed one) are **linearly dependent** which means that **this vector can be expressed as a linear combination of the others** (since it lies in the span of the others). On the other hand if each vector adds another dimension to the span there said to be linearly independent.

Linearly dependent

$$\vec{u} = a\vec{v} + b\vec{w}$$

For some values of a and b

linearly independent

$$\vec{w} \neq a\vec{v}$$

For all values of a

$$\vec{u} \neq a\vec{v} + b\vec{w}$$

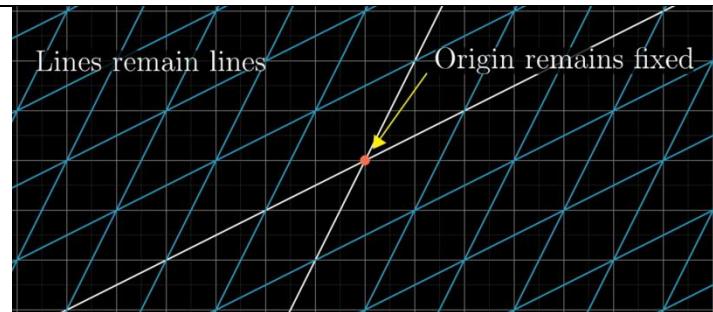
For all values of a and b

## Basis of a vector space

The basis of a vector space is a set of linearly independent vectors that span the full space

## Linear transformation

A linear transformation is a function that takes in vectors and spits out vectors. It can be represented visually by smooshing around space in such a way that grid lines remain parallel and evenly spaced and the origin remains fixed (if not evenly spaced the diagonals would not remain straight lines).



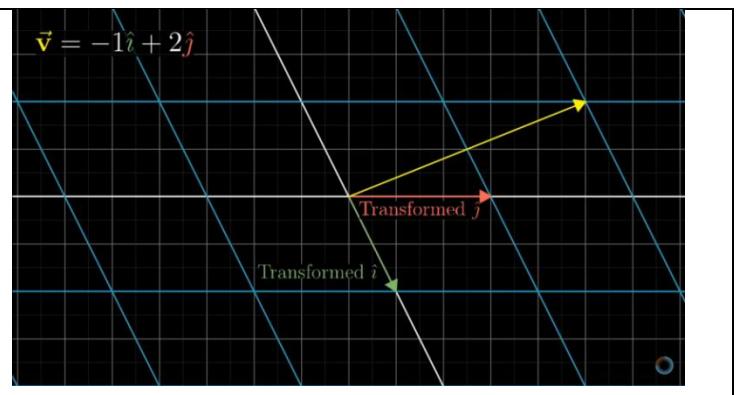
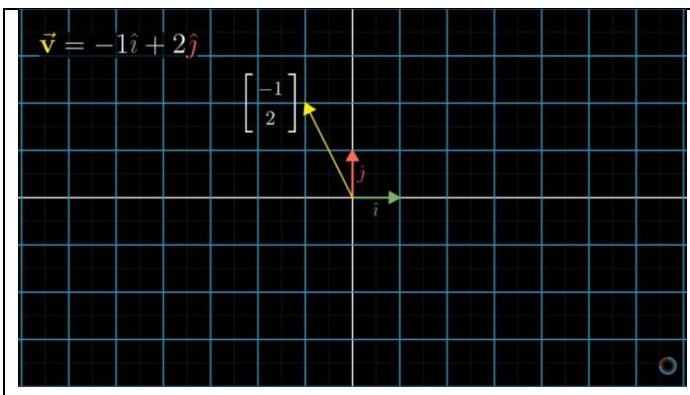
We can think of each point in a space as the tip of a vector. Then we draw on top (the cyan lines) the transformed vector (its coordinates) where it lands when it passes through a transformation (a function).

In a **linear transformation**, lines must remain lines and origin must remain fixed. Notice that diagonal lines too must remain straight lines. Or equivalently: **Grid lines must be kept parallel and evenly spaced and the origin fixed.**

We have space with its basis vectors and a vector  $v [-1, 2]$ . The vector  $v$  can be expressed as a linear combination of its space's basis vectors  $\vec{v} = -1\hat{i} + 2\hat{j}$ . We apply a linear transformation to the space and watch where the the  $v$  vector and the basis vectors land. **The fact that the grid lines are kept parallel and evenly spaced has an important consequence. The vector  $v$  started as a certain linear combination of the basis vectors and ended up as the same linear combination of the transformed basis vectors.** This means that you can deduce where every vector lands based solely on where the basis vectors land.

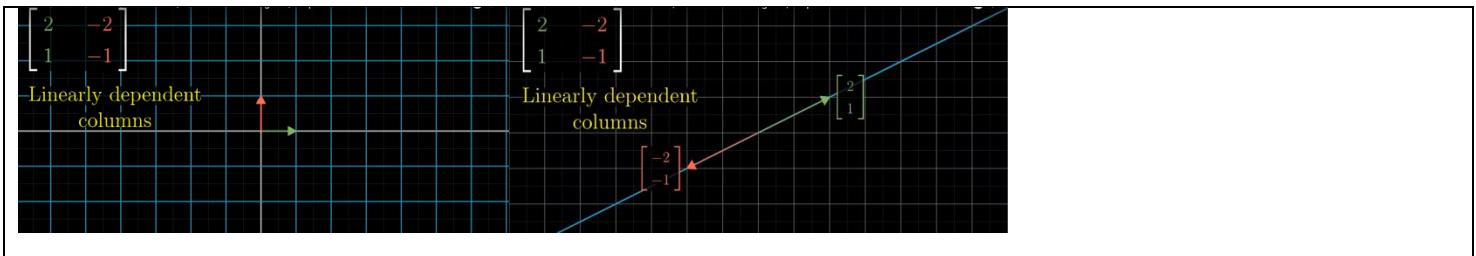
$$\vec{v} = -1\hat{i} + 2\hat{j}$$

$$\text{Transformed } \vec{v} = -1(\text{Transformed } i) + 2(\text{Transformed } j)$$



How would you describe a linear transformation numerically? What formula would you use so that the coordinates of each vector are transformed to the coordinates of the output vector? It is enough to know where the basis vectors land. That is enough to completely describe the transformation.

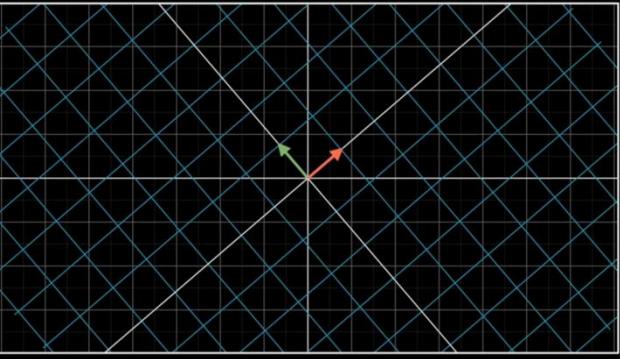
	<p>We can see that the x basis vector [i] lands in position [1,-2] (in terms of the original space) after the transformation and the y basis vector [j] lands in position [3,0].</p>
$\begin{aligned} &= -1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} + 2 \begin{bmatrix} 3 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -1(1) + 2(3) \\ -1(-2) + 2(0) \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \end{aligned}$	<p>Knowing the linear combination that gives the vector v we can numerically calculate the vector by addition. Notice that the resulting vector coordinates [5 2] are expressed in relation to the original basis vectors. The same vector has coordinates [-1 2] in relation to the new basis vectors which is actually the same linear combination of the transformed basis vectors.</p>
$\begin{aligned} \hat{i} &\rightarrow \begin{bmatrix} 1 \\ -2 \end{bmatrix} & \hat{j} &\rightarrow \begin{bmatrix} 3 \\ 0 \end{bmatrix} \\ \begin{bmatrix} x \\ y \end{bmatrix} &\rightarrow x \begin{bmatrix} 1 \\ -2 \end{bmatrix} + y \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 1x + 3y \\ -2x + 0y \end{bmatrix} \end{aligned}$	<p>So in general terms, <b>a vector with coordinates [x y] would land on x times where the [i] vector lands plus y times the vector of where the [j] lands.</b></p> <p>This formula can be represented by a vector-matrix multiplication.</p>
<p>“2x2 Matrix”</p> $\begin{bmatrix} (3) & (2) \\ (-2) & (1) \end{bmatrix}$ <p>Where <math>\hat{i}</math> lands      Where <math>\hat{j}</math> lands</p>	<p>This two by two matrix describes a two dimensional linear transformation and is composed of two columns, one for each basis vector (space dimension) where the columns describe where the basis vectors of the original space land if this transformation is applied.</p>
$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x \underbrace{\begin{bmatrix} a \\ c \end{bmatrix}}_{\text{Where all the intuition is}} + y \underbrace{\begin{bmatrix} b \\ d \end{bmatrix}}_{\text{Where all the intuition is}} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$	<p>If you are given any vector in this space you can easily deduce where it would land by multiplying its x coordinate with transformed [i] and y coordinate with transformed [j] and add. This corresponds to the idea of adding the new transformed basis vectors. This formula is equivalent of multiplying the vector with the matrix.</p>
<p>If you rotate -90 degrees around z then <math>i[1,0]</math> goes to <math>[0,1]</math> and <math>j[0,1]</math> goes to <math>[-1,0]</math>. So the transformation matrix would be:</p>	
$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$	<p>If the transformed basis vectors are linearly dependent (one is a scaled version of the other) then the linear transformation squeezes all of 2d space onto the line where these two vectors sit (the one dimensional span of these linearly dependent vectors).</p>



Every time you see a matrix you can interpret it as a certain transformation of space.

## Orthonormal transformations

The word "orthonormal" typically describes a set of vectors which are all unit length and orthogonal. (Orthogonal means that two things are 90 degrees from each other. Orthonormal means they are orthogonal and they have "Unit Length" or length 1.)

<p>If <math>T(\vec{v}) \cdot T(\vec{w}) = \vec{v} \cdot \vec{w}</math> for all <math>\vec{v}</math> and <math>\vec{w}</math>  <math>T</math> is "Orthonormal"</p> 	<p><u>They are linear transformations that leave all of the basis vectors perpendicular to each other and still with unit lengths.</u></p> <p>We often think of them as the rotation transformations. They correspond to <b>rigid body motion</b> (no stretching, squishing or morphing of the space)</p> <p>Since any vectors keep the same length and angle in between them before and after the transformation, their dot product remains the same.</p>
--	--

## Orthogonal matrices

Orthogonal matrices are the best to calculate with. They don't change the length of things, they don't result in blow up, they don't transform anything to 0. You can multiply a lot of them and you would still have an orthogonal matrix as a result.

<p>Orthogonal Vectors – Matrices – Subspaces</p> $x^T y = 0 \quad y^T x = 0 \quad (x + y)^T (x + y) = x^T x + y^T y \quad \text{RIGHT TRIANGLE}$ <p>Orthonormal columns <math>q_1, \dots, q_n</math> of <math>Q</math>: Orthogonal unit vectors</p> $Q^T Q = \begin{bmatrix} \cdots & q_1^T & \cdots \\ \vdots & & \ddots \\ \cdots & q_n^T & \cdots \end{bmatrix} \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{bmatrix} = I_n$ $Q Q^T = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} \begin{bmatrix} q_1^T \\ \vdots \\ q_n^T \end{bmatrix} = q_1 q_1^T + \cdots + q_n q_n^T = I$	<p>Orthogonal matrices are square matrices that have orthonormal columns. This means that the dot product of any two columns is 0. Because the dot product of a vector with itself is 1, <math>Q^T Q</math> and <math>Q Q^T</math> is the identity matrix. <math>Q^T Q = I</math> so <math>Q^T = Q^{-1}</math></p> <p>All columns are orthonormal with each other, which means that they are linearly independent with each other. So each column is a unit vector (that's why they are called orthonormal).</p> <p>The eigenvectors of an orthogonal matrix are all orthogonal to each other. This is another way of saying that the transformation is defined only by rotations where the axes of ratios are the</p>
---	--

## "Orthogonal matrix"

$Q = \frac{1}{3} \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}$  is square. Then  $QQ^T = I$  and  $Q^T = Q^{-1}$

If  $Q_1, Q_2$  are orthogonal matrices, so are  $Q_1Q_2$  and  $Q_2Q_1$

$$\|Q\mathbf{x}\|^2 = \mathbf{x}^T Q^T Q \mathbf{x} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2 \quad \text{Length is preserved}$$

$$\text{Eigenvalues of } Q \quad Q\mathbf{x} = \lambda\mathbf{x} \quad \|Q\mathbf{x}\|^2 = |\lambda|^2 \|\mathbf{x}\|^2 \quad |\lambda|^2 = 1$$

$$\text{Rotation } Q = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad \lambda_1 = \cos \theta + i \sin \theta \quad |\lambda_1|^2 = |\lambda_2|^2 = 1$$

## Gram-Schmidt Orthogonalize the columns of $A$

$$A = QR \quad Q^T A = R \quad \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} = \begin{bmatrix} q_1 & \dots & q_n \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{nn} \end{bmatrix}$$

Columns  $a_1$  to  $a_n$  are independent Columns  $q_1$  to  $q_n$  are orthonormal!

eigenvectors, which by itself defines a rigid body motion.

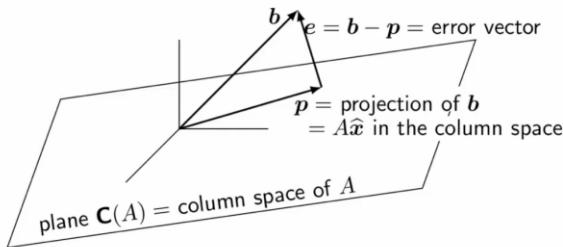
Rotations are orthogonal matrices.

## Least Squares : Major Applications of $A = QR$

$m > n$   $m$  equations  $A\mathbf{x} = \mathbf{b}$ ,  $n$  unknowns, minimize  $\|\mathbf{b} - A\mathbf{x}\|^2 = \|\mathbf{e}\|^2$

Normal equations for the best  $\hat{\mathbf{x}}$  :  $A^T e = 0$  or  $A^T A \hat{\mathbf{x}} = A^T b$

If  $A = QR$  then  $R^T Q^T Q R \hat{\mathbf{x}} = R^T Q^T b$  leads to  $R \hat{\mathbf{x}} = Q^T b$



BECAUSE ORTHOGONAL MATRICES HAVE THESE NICE PROPERTIES WE WANT TO DECOMPOSE ANY NON orthogonal matrix (with independent columns but not orthogonal) to an orthogonal and a triangular one.  $\mathbf{A}=\mathbf{QR}$

## Composition of Transformations

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}} \underbrace{\left( \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Rotation}} \begin{bmatrix} x \\ y \end{bmatrix} \right)}_{\text{Composition}} = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Composition}} \begin{bmatrix} x \\ y \end{bmatrix}$$

If you apply two linear transformations one after the other (for example rotation by 90 degrees first and a shear transformation afterwards) the resulting transformation is a new one and is called a composition of the other two. Since applying a transformation to a vector can be numerically represented with multiplying it by a matrix, the composition can be represented numerically with a matrix-matrix multiplication. Whenever you see such a multiplication it can be graphically represented as two transformations one after the other.

$$f(g(x))$$

Read right to left

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}} \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Rotation}} = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Composition}}$$

This stems from function notation where we write variables on the right so when you read a function composition you read it right to left

Notice that if you reverse the order of transformations the resulting one will be different. So in its numerical representation which is matrix multiplication order is important.

$M_1 M_2 \neq M_2 M_1$  Matrix multiplication is not cumulative

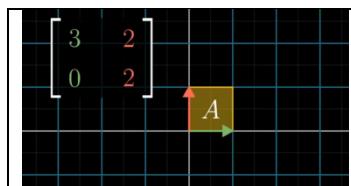
So, if transformations are applied in the same order give the same result, matrix multiplication is associative since in both cases all you do is apply all the transformations right to left.

$$(AB)C = A(BC)$$

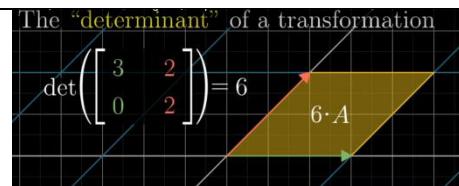
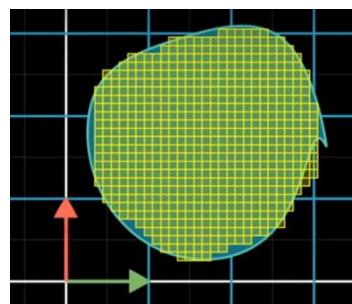
## The determinant

A linear transformation changes any area within its space by a certain factor. This scaling factor is called the determinant of the linear transformation. It has a sign plus or minus depending on the orientation of the transformed area. In more general terms, the determinant of a matrix shows how much the **signed unit object** changes after applying the transformation described by the matrix. If the space is 2d the unit object is an area, if it is a 3d space then it is a volume.

If you have a matrix you can think of the determinant of that matrix as the area formed between two vectors where each vector is a column of the matrix. But when the columns of a matrix represent vectors, the matrix represents a linear transformation where each column is a transformed basis vector. The basis vectors form the unit object. The initial unit object with an area of size one has been transformed to a new unit object with a new area the size of which is the determinant value. Thinking in terms of unit objects, we can say that the determinant is the factor by which the initial unit object changed. Since any object can be expressed as a sum of unit objects the determinant is the factor by which the volume of any object changes.



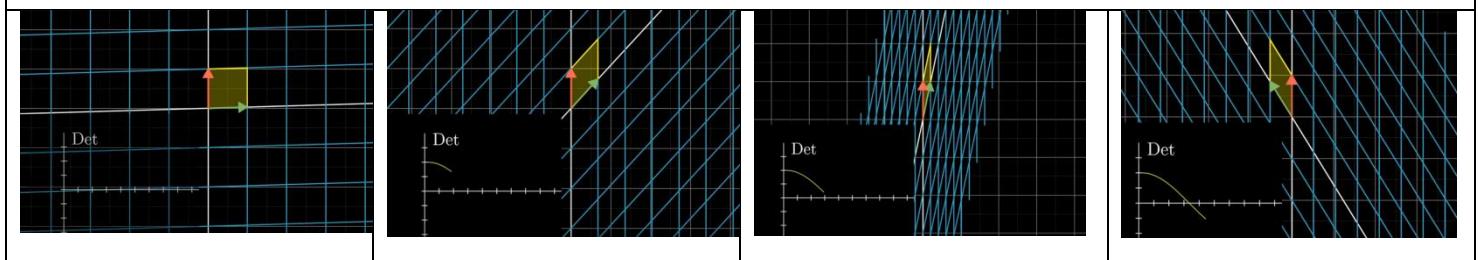
Any shape that is not a grid cell can be approximated well enough if you use small enough grid cells, so its change in size can be approximated by the change in size of the grid cells.



The determinant of a transformation is **zero** if it squishes all space onto a line or a point in case of 2d space, or onto a 2d plane, line or point in case of 3d space. In general if the determinant of a transformation is 0, it means that this transformation **squishes space into a smaller dimension** (3d to 2d or 1d or 0d). This happens when the transformed basis vectors (the columns of the matrix) become linearly dependent.

(A singular matrix refers to a matrix whose determinant is zero. Furthermore, such a matrix has no inverse.)

	$\det\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} = 0$
$\det\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix} = -3.0$	The determinant can be a <b>negative</b> number. This has to do with orientation. If a transformation flips space over ( <b>inverts the orientation of space</b> ) then the area is represented with a negative number.



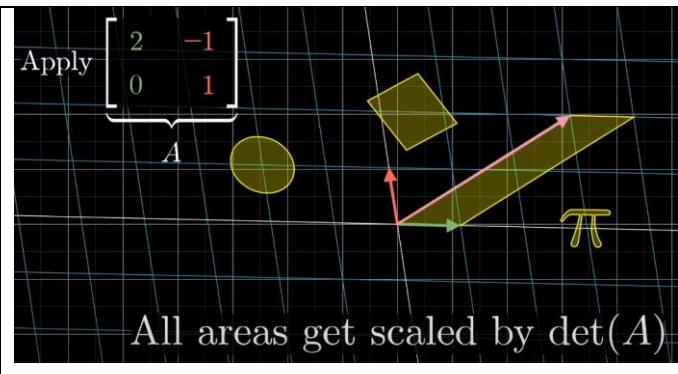
In 3d space the determinant of a  $3 \times 3$  matrix (each column is the coordinates of a vector) shows the factor by which a volume changes if the linear transformation represented by this matrix is applied. In 3d space you can determine if the space orientation has been inverted by the right hand rule. If it makes sense to represent the orientation with your left hand instead, it means that the space has been inverted (bring the in of a 3d object, out)

The numerical formula for computing the determinant comes from computing the area of the transformed unit object. For a linear transformation of a 2d space which can be represented numerically by a  $2 \times 2$  matrix the determinant is  $ad - bc$ .

$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$	$\det\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a \det\begin{pmatrix} e & f \\ h & i \end{pmatrix} - b \det\begin{pmatrix} d & f \\ g & i \end{pmatrix} + c \det\begin{pmatrix} d & e \\ g & h \end{pmatrix}$
	$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (a+b)(c+d) - ac - bd - 2bc = ad - bc$

Also

$$\det(M_1 M_2) = \det(M_1) \det(M_2)$$



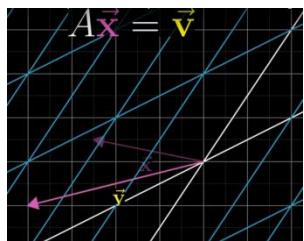
Have in mind this key idea about determinants. All shapes within the initial space are scaled by the same amount (that amount is the determinant of the transformation)

## Inverse matrix

Usefulness of linear algebra

- Describing the manipulation of space (useful for computer graphics and robotics)
- It helps solve systems of linear equations (useful for a broad range of technical areas)

$$\begin{array}{l} 2x + 5y + 3z = -3 \\ 4x + 0y + 8z = 0 \\ 1x + 3y + 0z = 2 \end{array} \rightarrow \underbrace{\begin{bmatrix} 2 & 5 & 3 \\ 4 & 0 & 8 \\ 1 & 3 & 0 \end{bmatrix}}_A \underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix}}_{\vec{v}}$$



A linear system of equations: The only thing that happens to a variable is that it is scaled by a constant and the only thing that happens to those scaled variables is that they are added to each other. Any such system can be represented by a vector-matrix multiplication and as such it can be thought of as a linear transformation of a space. (Notice that this chapter only mentions systems which have the same number of equations and unknowns)

$$\det(A) \neq 0 \quad A^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The transformation  $A^{-1}$  exists that does nothing

$$A^{-1}A\vec{x} = A^{-1}\vec{v}$$

You know  $A$  which is the linear transformation applied, you know the transformed vector, so what you have to find in this case is the inverse transformation that transforms the transformed vector back to the original one. The inverse transformation is represented by a matrix  $A^{-1}$  which is the inverse matrix of  $A$ .

Applying two transformations one after the other is numerically represented by matrix multiplication so  $A \cdot A^{-1}$  results to a transformation that does nothing and is called the Identity transformation.

If you know the inverse, you multiply both terms of the equation with it and you end up with the solution which is applying the inverse transformation to  $v$  in order to get  $x$ .

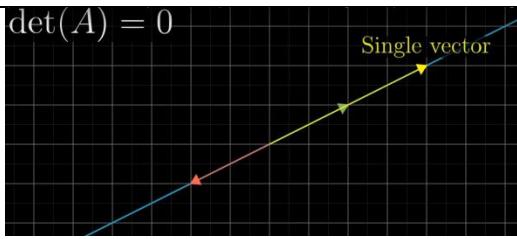
It is important to separate two cases:

- The determinant of  $A$  is non zero

A matrix  $A^{-1}$  exists with property that when you do  $A$  and then you do  $A^{-1}$ , it's the same as doing nothing.

- The Determinant is zero

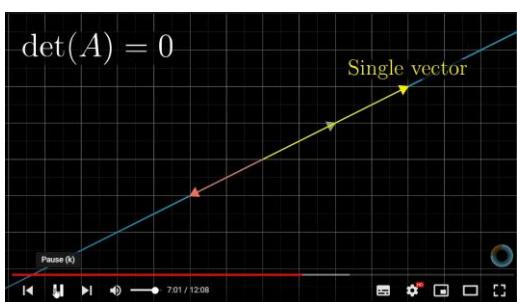
$$\det(A) = 0$$



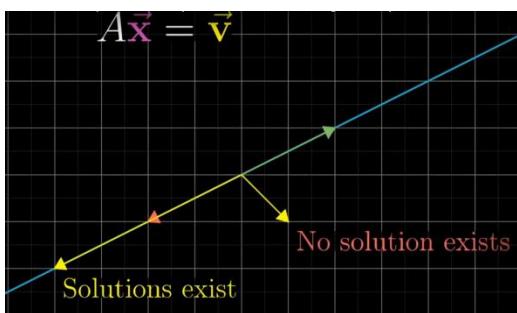
Determinant is zero

If the determinant is zero then there is no inverse transformation which means that there is no solution to that system of linear equations (except if the transformed vector  $v$  lies along the line).

$$\det(A) = 0$$



$$A\vec{x} = \vec{v}$$



No solution exists

Solutions exist

There is no transformation (no function) that can be applied to a line to transform it to a plane. This is not something that a function can do. If there was such a function, it would mean that it transforms an individual 1d vector into a line of vectors in the 2d space (the slightly larger 1d vector would be transformed to another line right next to the previous one and so on, so that all 1d vectors are transformed to infinite number of lines that form a 2d space). This means that all the vectors that lie in this line of the original space, are transformed to a single vector (a point) in the transformed space (which is a line). A function by definition maps a single input to a single output. So there is no function that can map a single vector to the infinite vectors that form a line. Although the opposite can be done by a function. It maps many inputs to the same output. This is the original linear transformation.

Notice that a solution to the system could still exist but this is only in the case the transformed vector  $v$  lies within the transformed space (along the line in this case), in which case the solution is all vectors that belong to a specific line in the original space.

If the vector  $v$  lies outside of the line, it lies outside of the space which was defined after the transformation. Thus it can't be a result of that transformation. Every vector that was lying inside the original space, would lie after the transformation inside the transformed space which is the line. If instead the transformed vector lives in the line then there could be a vector  $x$  that when the linear transformation is applied to it, it ends up in  $v$ . (in other words if the vector  $v$  lives in the column space then a solution to the equation exists). Actually there are a lot of vectors that land on  $v$ . A line of vectors, which is composed of all vectors the tip of whom lies in a specific line, are the solution to this equation.

## Conditioning and stability

<https://www.youtube.com/watch?v=BVM3NAt6QoM>

Conditioning shows how error is magnified in computational problems.

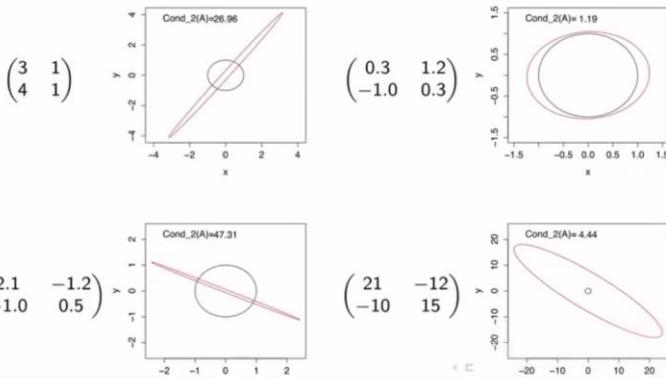
- ▶ **Conditioning** refers to the perturbation behavior of a problem
  - ▶ Problem is well-conditioned if a small perturbation in the input data leads to a small perturbation in the output of the problem
  - ▶ Problem is ill-conditioned if a small perturbation in the input data leads to a large perturbation in the output of the problem
- ▶ **Stability** refers to the perturbation behavior of a numerical algorithm to solve that problem on a computer

In computational (numerical) methods you approximate the exact solution. The conditioning of the problem can tell you if your solution is a good approximation or not.

## Matrix condition number

### Examples of the Condition Number for 2x2 Matrices

- ▶ Image of the unit sphere in the 2-norm under a linear mapping  $A$  is a hyperellipse
- ▶ Using the 2-norm, the condition number of  $A$ ,  $\kappa_2(A)$  gives the ratio of the length of the longest principal semiaxis to the length of the shortest principal semiaxis (maximum and minimum singular values)



If the condition number is **not too much larger than 1**, the matrix is well-conditioned, which means that its inverse can be computed with good accuracy (with a numerical method).

The condition number of a matrix  $n$  by  $n$  is calculated by multiplying the norm of  $A$  by the norm of  $A^{-1}$ . The norm (of all rows of the matrix) is the largest sum of all rows (you sum each row. The largest sum is the norm of  $A$ ). you can think of this expression like this. A matrix applies a transformation to a space. Some parts of the space are expanded more than others. The ratio of the maximum expansion with the minimum expansion is the condition number.

## Error magnification

### Error Magnification in Solving $A\vec{x} = \mathbf{b}$

Suppose we solve  $A\vec{x} = \mathbf{b}$  and get an approximate answer  $\vec{x}_a$

	Backward Error	Forward Error
Absolute	$\ \mathbf{b} - A\vec{x}_a\ $	$\ \vec{x} - \vec{x}_a\ $
Relative	$\frac{\ \mathbf{b} - A\vec{x}_a\ }{\ \mathbf{b}\ }$	$\frac{\ \vec{x} - \vec{x}_a\ }{\ \vec{x}\ }$

The **error magnification** is the number given by

$$(\text{relative backward error}) \times (\text{error magnification}) = (\text{relative forward error})$$

or

$$(\text{error magnification}) = \frac{(\text{relative forward error})}{(\text{relative backward error})} = \frac{\left( \frac{\|\vec{x} - \vec{x}_a\|}{\|\vec{x}\|} \right)}{\left( \frac{\|\mathbf{b} - A\vec{x}_a\|}{\|\mathbf{b}\|} \right)}$$

You can calculate how good approximation is a solution. You calculate the magnification error of the solution. If it is small, then it means that your approximation happened to be a good one.

The condition number bounds the error magnification (as it is called). The error magnification of any algorithm will be smaller than the condition number of the matrix. The condition number kind of shows the worst-case scenario.

Have in mind the forward error and the backward error. Fwd is the (relative) difference between the approximated  $x$  you found and the exact one. The backward error is the (relative) difference between the  $b$  (the result of the exact  $x$ ) and the  $Ax_1$  (the  $b$  you calculate with the approximated solution)

This is an important take away

In an ill conditioned problem, a small residual doesn't necessarily mean that you found a good solution to the problem. In an ill conditioned  $Ax=b$  problem, the  $x$  you numerically approximated might be very different from the exact one, but they can both result on almost the same  $y$ .

## Transpose matrix

PROVE:  $A^T A$  is square, symmetric, nonnegative definite

1.  $A^T A = (n \times m)(m \times n) = n \times n$  Square

2.  $(BA)^T = A^T B^T$      $(A^T A)^T = A^T A^{TT} = A^T A$  Symmetric

3.  $S = S^T$  is nonnegative definite IF

EIGENVALUE TEST 1: All eigenvalues  $\geq 0$      $Sx = \lambda x$

ENERGY TEST 2:     $x^T S x \geq 0$  for every vector  $x$

$A^T A$  is square, symmetric and nonnegative definite (its eigenvalues are non negative)

## Rank, Column space and Null space or kernel

Rank is the number of dimensions of the output space of a transformation. When the result of a linear transformation is a line (which has one dimension) we say that the transformation has a rank of one. If the space is transformed to a 2d space we say that it has a rank of 2. The columns of a matrix are the coordinates of the transformed basis vectors. The span of those basis vectors gives you all the possible outputs of the transformation and is called **column space**. In these terms, the **rank is the number of dimensions of the column space of a matrix**. If the rank is the highest it can be, in other words if it is the same with the number of columns of a matrix, we say that the matrix is **full rank** and the linear transformation that it describes transforms a space into another space of the same number of dimensions. Notice that the **zero vector** will always be included in the column space of any matrix, since it lies in the origin and linear transformations must keep the origin fixed in space (the origin is the lowest possible result of a transformation so the origin is always part of the result).

In a full rank linear transformation the only vector that ends up in the origin is the zero vector. In an one rank transformation of a 2d space (which is represented by a  $2 \times 2$  matrix) there is a line in the original 2d space all the vectors of which land in the origin after the transformation. If a 3d transformation squishes space into 2 dimensions, a line exist in the 3d space the vectors of which (a line of vectors) land in the origin. If a 3d transformation squishes space into one dimension, a 2d plane exists in the original space all the vectors of which land on the origin of the transformed space. This set of vectors that land on the origin after a linear transformation is called the **null space** or the **kernel** of the matrix. It is the space of all vectors that become null.

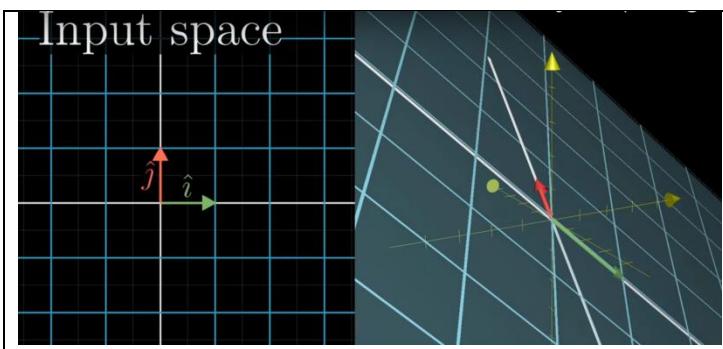
	<p>The kernel of a 2d linear transformation with rank one is a line (the yellow line). The transformed space is a line too (the cyan line).</p>
$A \vec{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	<p>In case of a system of linear equations, if the result vector is the zero vector then the null space is all the possible solutions to the equation.</p>

Have in mind these terms for how to numerically calculate the solutions to systems of linear equations:

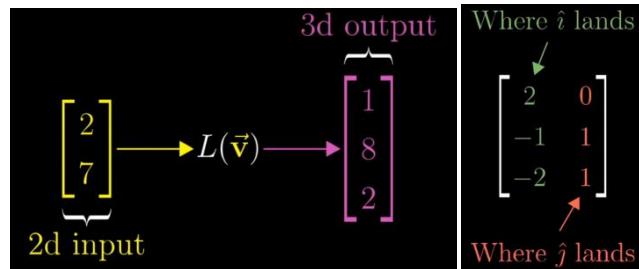
- Gaussian elimination
- Row echelon form

## Non square matrices

Non square matrices describe transformations between dimensions.

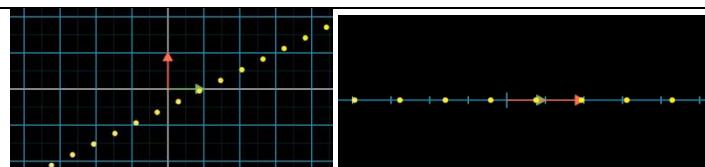


A 3 by 2 matrix has the geometric interpretation of transforming a 2d space to a 3d space. Notice that a 3\*2 matrix is a full rank matrix which means that it transforms space into a space of same number of dimensions (but as part of a space with one more dimension). It has two columns which means that the output space is a 2d plane and the transformed basis vectors have 3 coordinates which means that they lie in a 3d space. But the vectors that are transformed with a 3\*2 matrix will always lie in the same 2d plane.



Since each column show the where the basis vector of the original space lands (the transformed basis vector), the number of columns show the number of basis vectors so the number of dimensions of the original space.

The number of rows show the number of coordinates of the basis vectors in the transformed output space.



$$\text{Transformation matrix: } \begin{bmatrix} 2 & 1 \end{bmatrix}$$

$\hat{i}$  lands on 2  
 $\hat{j}$  lands on 1

Number line from -4 to 5

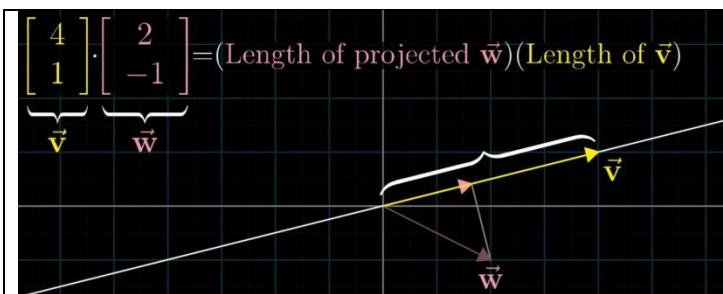
You can transform a 2d space onto a 1d space (with a 1\*2 matrix). It takes in vectors and outputs numbers. When space is squished onto a line it is difficult to think of grid lines that remain at equal distance. A way to visualize linearity in this case is to think of a line of evenly spaced dots that would remain evenly spaced as they are mapped onto the line.

Since the transformed space is a line each basis vector would be represented by a single number.

Notice that a transformation  $[2 \ 1]$  is different from another let's say  $[2 \ -1]$ . it transforms the space in a different way so each vector lands on different numbers in each case.

## Dot product

The dot product is a useful geometric tool for understanding projections and determining if two vectors are in the same or opposing direction or perpendicular to each other.



The dot product of two vectors can be geometrically represented by finding the projection of one vector onto the other and multiplying the projected length with the vector. The result would be the same no matter which vector you select to project.

Vertical vectors -> dot product is 0

Vectors with opposing direction -> dot product is negative

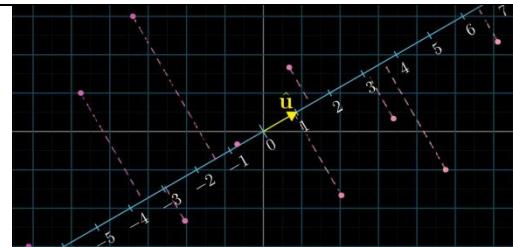
$$\begin{bmatrix} 2 \\ 7 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 2 \\ 8 \end{bmatrix} = 2 \cdot 8 + 7 \cdot 2 + 1 \cdot 8$$

Dot product

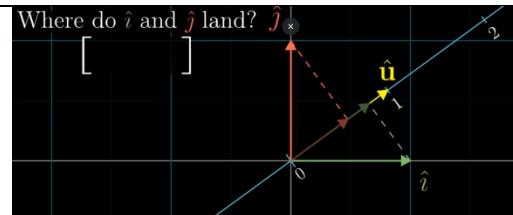
Order doesn't matter

Why the formula of dot product has anything to do with projections?

When you want to see where a vector  $[4 3]$  lands after a  $[1 2]$  transformation ( $[1 2]$  is a matrix) you can think that  $[4 3]$  is  $4i+3j'$  and since the result of a linear transformation would be the same linear combination of the transformed basis vectors  $4i'+3j'$  you just have to do  $4*1+3*2=-2$ . This is the number where the 2d vector will land. Numerically this formula is a matrix vector multiplication. But if you write the  $[1 2]$  matrix as a vector then the dot product formula of the two vectors  $[1 2]$  and  $[4 3]$  gives the same result as with the equivalent matrix vector multiplication. So it seems that there is some kind of association between linear transformations that map 2d spaces to lines (represented by a  $1*2$  matrix) and vectors of size two themselves.



Imagine that you place a number line inside a 2d space so that it passes from the origin and forms an angle with the x axis of the 2d space. You can draw a vector  $u$ , which has the same length with the unit vector  $i$  of the 2d space but it lies in the same direction with the number line. We mark the number one of the number line in the length of the  $u$ . Now, if we take some random 2d vectors we can just project them onto this number line. By definition they are transformed to numbers. What we have done actually, is that we have defined a function that takes 2d vectors and outputs numbers and this function is linear since it keeps the dots (tips of vectors) evenly spaced in the output space (the line). So this projection is actually a linear transformation from the 2d plane to this specific line. Consequently we can find a  $1*2$  matrix that describes this transformation.



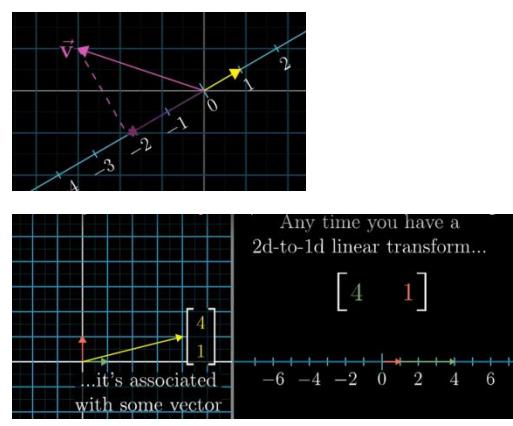
To do so we have to find where the  $i$  and  $j$  basis vectors of the 2d plane land on the line. They land on their projections on the  $u$  vector. But since the  $u$  and  $i$  vectors are both unit vectors (of length one) the projection of  $i$  on  $u$  is of the same length with the projection of  $u$  to  $i$ . The projection of  $u$  to  $i$  is the x coordinate of  $u$  ( $ux$ ). Respectively the projection of  $j$  in  $u$  is of length ( $uy$ ). So the basis vectors land on  $ux$  and  $uy$  which are numbers of the number line. This way we have found the  $1*2$  matrix that describes this linear transformation. It is the  $[ux \ uy]$  matrix.

$$\begin{bmatrix} u_x & u_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = u_x \cdot x + u_y \cdot y$$

Matrix-vector product  
↓  
Dot product

$$\begin{bmatrix} u_x \\ u_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = u_x \cdot x + u_y \cdot y$$

This means that when you want to transform a vector to this line you multiply its coordinates with  $ux$  and  $uy$  (the  $u$  matrix) and add them together. But this formula is identical with the formula of the dot product between the vector and the  $u$  vector.



Consequently, taking the dot product of a vector with a unit vector can be interpreted as projecting the vector onto the span of this unit vector and taking the projection's length (since the length of the unit vector is one and multiplying a number by one is the same number, the dot product is the length of the projection)

The dot product with a non unit vector (following the same logic) is multiplying the projections length with the length of the non unit vector (which is a number times the unit vector)

## Duality

## Duality $\Leftrightarrow$ Natural-but-surprising correspondence

In general a 2d linear transformation to 1d is fundamentally related with a 2d vector. This is an example of something in math that is called “duality”, when you have a natural but surprising correspondence between two things.

- The dual of a vector is the linear transformation that it encodes
- The dual of a linear transformation from some space to one dimension (a line) is a certain vector in the original space.

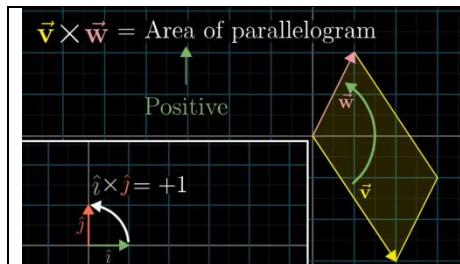
This means that **we can think of a vector not only as an arrow in space but as a certain transformation that transforms its space to a line**. It is as if the vector is a conceptual shorthand for a certain transformation

Duality applies to transformations from a space of any number of dimensions to a space of one dimension

## Cross product

Note, in all the computations here, I list the coordinates of the vectors as columns of a matrix, but many textbooks put them in the rows of a matrix instead. It makes no difference for the result, since the determinant is unchanged after a transpose, but given how I've framed most of this series I think it is more intuitive to go with a column-centric approach.

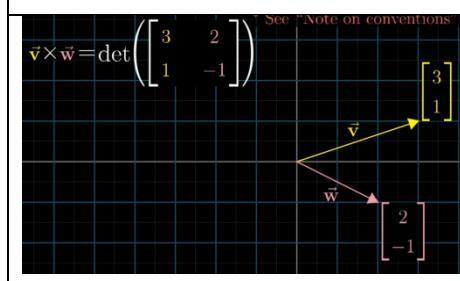
A 2d cross product (of two 2d vectors) is the area between the two vectors with a s=plus or minus sign depending on the orientation.



The cross product can take negative values too depending on orientation . In the cross product  $v^*w$ , if  $v$  (the first vector) is on the right of  $w$  (the second vector) the result is positive.

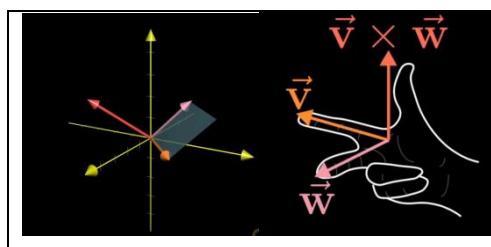
$$\vec{v} \times \vec{w} = -\vec{w} \times \vec{v}$$

$$(3\vec{v}) \times \vec{w} = 3(\vec{v} \times \vec{w})$$



The cross product of two vectors is the determinant of a matrix whose columns are the vectors coordinates. That is because that matrix corresponds to a linear transformation that transforms the perpendicular basis vectors  $i$  and  $j$  of the original space to the  $v$  and  $w$  vectors respectively. The determinant shows how much an area changes after the transformation but in this case the initial area formed by the two initial basis vectors is one. This means that the determinant would be equal to the final area, which is the cross product. If vectors are perpendicular the size of the area that they form is bigger than if they were close to parallel. If you scale one vector by a factor, the area that it forms with another vector also scales by the same factor.

The “true” cross product, the 3d cross product (of two 3d vectors) is not a number, it is a vector the magnitude of which is the area between the two vectors that is perpendicular to the plane defined by the two vectors with a direction defined by the right hand rule.



The vectors length is the area of the parallelogram and its direction is perpendicular to the plane the two vectors form. The direction of the vector is extracted from the right hand rule (from the orientation of the basis vectors of the plane)

The formula for the cross product of two 3d vectors can be extracted if we take the determinant of a  $3 \times 3$  matrix where the first column is just

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \det \left( \begin{bmatrix} \hat{i} & v_1 & w_1 \\ \hat{j} & v_2 & w_2 \\ \hat{k} & v_3 & w_3 \end{bmatrix} \right)$$

$\underbrace{\hat{i}(v_2w_3 - v_3w_2)}_{\text{Some number}} + \underbrace{\hat{j}(v_3w_1 - v_1w_3)}_{\text{Some number}} + \underbrace{\hat{k}(v_1w_2 - v_2w_1)}_{\text{Some number}}$

the basis vectors of the 3d space. It has no meaning to put a vector in place of a number in a matrix, but it is just a notation thing in order to end up with a linear combination of those basis vectors which gives the cross product vector.

Facts you could (painfully) verify computationally
$\vec{v} \cdot (\vec{v} \times \vec{w}) = 0$
$\vec{w} \cdot (\vec{v} \times \vec{w}) = 0$
$\theta = \cos^{-1}(\vec{v} \cdot \vec{w} / (\ \vec{v}\  \cdot \ \vec{w}\ ))$
$\ (\vec{v} \times \vec{w})\  = (\ \vec{v}\ )(\ \vec{w}\ ) \sin(\theta)$

Understanding cross product in terms of linear transformations (the proof of the above formulas)

Not the real cross product  $\times$

$$\vec{u} \times \vec{v} \times \vec{w} = \det \left( \underbrace{\begin{bmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{bmatrix}}_{\text{Number}} \right)$$

Imagine that from the 2d cross product equivalent, you have to assume what a 3d cross product would look like. Intuitively you could think of three vectors which you combine into a matrix and get its determinant which corresponds to the volume between the three vectors and a plus or minus depending on the orientation. The determinant is a number not a vector, but we know that the cross product is a vector.

$$f \left( \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) = \det \left( \begin{bmatrix} x & \vec{v} & \vec{w} \\ y & v_1 & w_1 \\ z & v_2 & w_2 \\ & v_3 & w_3 \end{bmatrix} \right)$$

Variable

**The determinant of a 3x3 matrix is equivalent to a function that gets in three vectors and outputs a number or in other words a linear transformation from 3d to 1d space.** So there is a way to describe this function as matrix-vector multiplication. Assume that we replace the first vector with a variable vector (three variables for each coordinate) so that now the function is a function of three variables. Due to the duality the matrix-vector multiplication can be represented by an equivalent vector-vector dot product. So we are looking to find this dual vector p.

$$\begin{bmatrix} ? & ? & ? \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det \left( \begin{bmatrix} x & \vec{v} & \vec{w} \\ y & v_1 & w_1 \\ z & v_2 & w_2 \\ & v_3 & w_3 \end{bmatrix} \right)$$

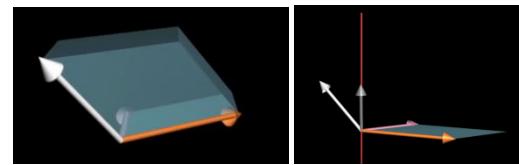
1 × 3 matrix encoding the 3d-to-1d linear transformation

What vector p has the property that when taking its dot product with any random vector [x y z] has the same result as plugging x y z as the first column of a matrix with fixed 2<sup>nd</sup> and 3<sup>rd</sup> columns as v and w and taking its determinant.

Or the geometrical equivalent

This is the numerical formula for the vector p that we are looking for that is the cross product of v and w.

what vector p has the property that when taking its dot product with a vector [x y z] it gives the same result as calculating the volume between the vectors v, w and [x y z]. xyz is shown here with white color.



$$\begin{bmatrix} \vec{p} \\ p_1 \\ p_2 \\ p_3 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det \begin{pmatrix} x & \vec{v}_1 & \vec{w}_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{pmatrix}$$

$$p_1 \cdot x + p_2 \cdot y + p_3 \cdot z = y(v_3 \cdot w_1 - v_1 \cdot w_3) + z(v_1 \cdot w_2 - v_2 \cdot w_1)$$

The volume of this object (parallelopiped) is the same with multiplying the area of the parallelogram with the component of xyz perpendicular to v and w plane.

But this is the definition of the dot product between xyz and a vector perpendicular to v and w with length equal to the area of the parallelogram  $v^*w$ ! So the vector that we are looking for is this vector and is the cross product of v and w.

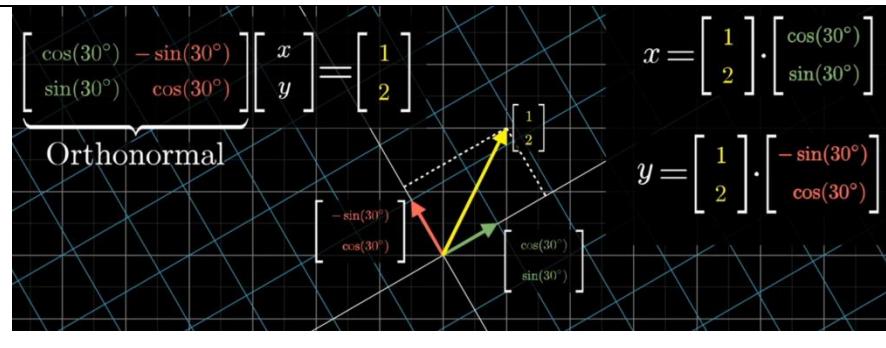
The formula that we found before for the cross product must correspond to a vector that is perpendicular to v and w and has a length equal to the area between v and w.

## Cramer's Rule

### Solving linear systems of equations

There are several computational methods of solving systems of linear equations using linear algebra. Cramer's rule is one of them. We will see the geometric interpretation of Cramer's Rule. Although have in mind that Gaussian Elimination is always a faster method.

### Systems with Orthonormal matrices

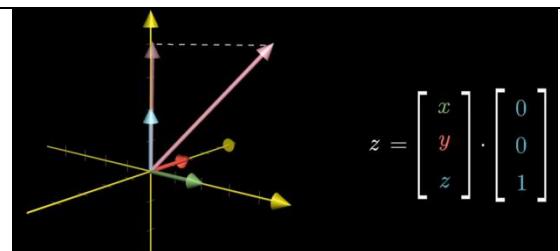


Solving linear systems of equations where the constants matrix represents an orthonormal transformation is very easy. The dot product of the transformed vector with the transformed i vector would be the same with the dot product of the original vector with the original i.

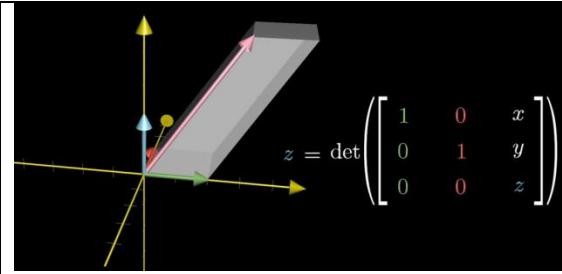
But the dot product of the original vector with i is the projection of vector onto i, which is the x value. So the dot product of the transformed vector with the transformed I vector (which are known) are the x and the dot product with the transformed j is y. You can find  $[x \ y]$  with two dot products.

### Non orthonormal matrices

We want to find the solution to a linear system of equations which is represented by a non orthonormal transformation  $A \cdot x = v$ . In a 3d space case, we want to find the x, y and z coordinates of vector x.

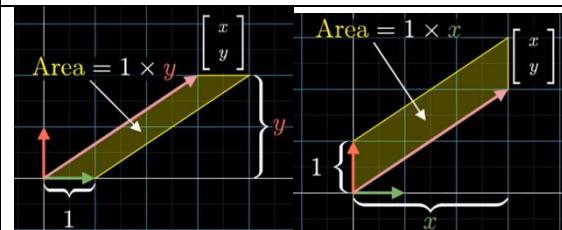


You can represent the coordinate of a vector in any of its basis vectors, with its dot product with this basis vector. Another way to represent it is with the signed volume of the parallelepiped that this vector forms with the other two basis vectors. The area of this object is one, so its volume which is its area times its height would be its height which is the coordinate of the vector in that axis. But this signed volume is the determinant of the three vectors that form it. Equivalently you can represent the other coordinates with the respective determinants. So we express the x,y and z unknown coordinates as determinants.

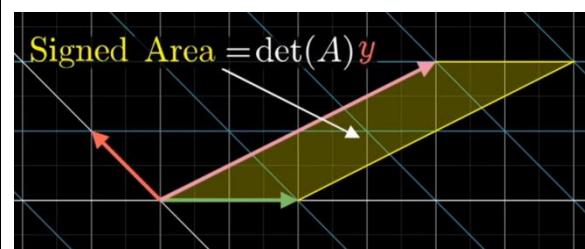


$$z = \det \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & z \end{pmatrix} \quad y = \det \begin{pmatrix} 1 & x & 0 \\ 0 & y & 0 \\ 0 & z & 1 \end{pmatrix} \quad x = \det \begin{pmatrix} x & 0 & 0 \\ y & 1 & 0 \\ z & 0 & 1 \end{pmatrix}$$

These volumes are scaled by the same amount after a transformation is applied to the space. They are scaled by a factor which is equal to the determinant of the transformation matrix.



$$\begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$



Using a 2d space for simplicity, we can represent x and y with the areas of the respective parallelograms, which is the respective determinants. After the transformation the area between the transformed i and transformed x (the vector v) is equal with the initial equivalent area times the scaling factor which is the determinant of the matrix. But the initial area is x so  $y = \text{Area}/\det(A)$

The Area is the determinant of a matrix the first column of which is the transformed basis vector I and the other is the result vector v. We do a similar process for x. This formula is the **Cramer's rule**. It can be generalized for more dimensions.

$$y = \frac{\text{Area}}{\det(A)} = \frac{\det \begin{pmatrix} 2 & 4 \\ 0 & 2 \end{pmatrix}}{\det \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}} \quad x = \frac{\text{Area}}{\det(A)} = \frac{\det \begin{pmatrix} 4 & -1 \\ 2 & 1 \end{pmatrix}}{\det \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}}$$

## Change of basis

Any way to transform a set of numbers to vectors is called a coordinate system. You can think of a vector's coordinates as scalars that scale the basis vectors of the space.

If you want to solve a linear system of equations  $Ax=v$  then the output vector v must be expressed relative to the initial coordinate system (the basis vectors of the space as it was before the transformation). If you know the output vector coordinates relative to the transformed coordinate system (the transformed basis vectors), then you have to transform it first to the initial basis and then solve the linear system.

## How to change basis

We know that a vector v is expressed in terms of a coordinate system as  $[ -1 \ 2 ]$ . We know the basis vectors of that coordinate system relative to ours  $b_1=[2 \ 1]$  and  $b_2=[-1 \ 1]$ . This means that we can think of this situation as a linear transformation applied to our space and transforming it to the new space. The transformation is represented by the known 2by2 matrix with columns the transformed basis vectors  $[2 \ 1]$  and  $[-1 \ 1]$  expressed in the initial basis. The vector v can be expressed as a linear combination of its basis vectors. Since we know this linear combination (the vector coordinates) and the basis vectors expressed in the initial basis, we can express the vector v in our coordinate system (the initial basis) by multiplying its x and y coordinates with the respective basis vectors. The result is the vector v expressed in our basis. This is a change of basis from the transformed basis to the initial basis and it is actually a vector matrix multiplication  $v_0=Av_1$ .

$$\begin{bmatrix} ? \\ ? \end{bmatrix} = -1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

Imagine that you have a vector  $x$  that lives in a certain space and a linear transformation ( $A$ ) is applied to that space to transform it to a new one  $\text{Space0} \rightarrow A_{01} \rightarrow \text{Space1}$ . The vector  $x$  is transformed along with the space and ends up in a new location. This can be represented by the typical linear system  $Ax=v$

- If you know the transformation matrix ( $A$ ) and the vector  $v$  expressed in the transformed space basis ( $v_1$ ) and you want to express the vector to the initial space basis ( $v_0$ ), you do a change of basis. You can think of this situation as a transformation applied to a vector ( $x$ ) that transformed it to  $v$ , but  $v$  is expressed in the transformed basis ( $Ax=v_1$ ). If you just want to find how  $v_1$  is expressed in the initial basis you just do  $v_0=A_{01}v_1$ .
- If the vector after a transformation is expressed in the initial space coordinates ( $v_0$ ) and you want to express it in the transformed space, the situation is expressed with the typical  $Ax=v_0$  linear system. In this case you have to compute the inverse of  $A$ .  $AA^{-1}x=A^{-1}v_0 \rightarrow x=A^{-1}v_0$ .  $x$  is actually the same linear combination both in the initial and in the output space, so  $x$  coordinates are actually the same coordinates with  $v$  as it is expressed in the transformed basis. So  $x=v_1$  and  $v_1=A^{-1}v_0$

You have a vector expressed in another basis. A known transformation (expressed in your basis) is applied to that space. The vector is transformed with the space and you want to express the transformed vector in its space basis (the beings of that space want to know where their vector would land after your transformation)

**Any vector in base 1 -> change of basis -> transformation in that basis -> inverse change of basis.** The formula for this can be written as  $(A_{01}^{-1}M_0A_{01}v_1)$  or in general terms:  $A^{-1}MAv$ . The  $A^{-1}MA$  notation (read right to left) is quite common so whenever you see it you can think in terms of **shift of perspective**.

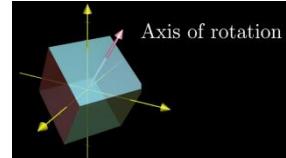
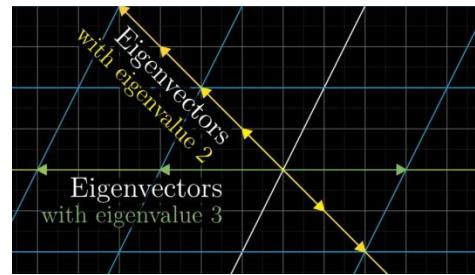
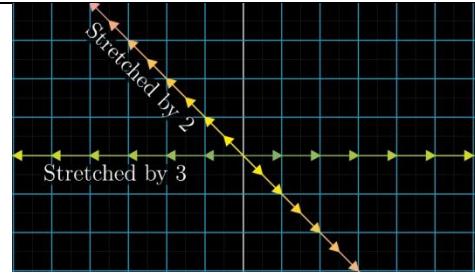
$A_{01}v_1=v_0$ <div style="border: 1px solid black; padding: 10px; background-color: black; color: white;"> <p>Same vector in our language</p> <math display="block">\underbrace{\begin{bmatrix} 2 &amp; -1 \\ 1 &amp; 1 \end{bmatrix}}_{\text{Change of basis matrix}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}</math> </div>	$M_0A_{01}v_1=v_0'$ <div style="border: 1px solid black; padding: 10px; background-color: black; color: white;"> <p>Same vector in our language</p> <math display="block">\underbrace{\begin{bmatrix} 0 &amp; -1 \\ 1 &amp; 0 \end{bmatrix}}_{\text{Transformation matrix in our language}} \underbrace{\begin{bmatrix} 2 &amp; -1 \\ 1 &amp; 1 \end{bmatrix}}_{\text{Change of basis matrix}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}</math> </div>	$A_{01}^{-1}M_0A_{01}v_1=v_1'$ <div style="border: 1px solid black; padding: 10px; background-color: black; color: white;"> <p>Transformed vector in our language</p> <math display="block">\underbrace{\begin{bmatrix} 2 &amp; -1 \\ 1 &amp; 1 \end{bmatrix}}_{\text{Inverse change of basis matrix}}^{-1} \underbrace{\begin{bmatrix} 0 &amp; -1 \\ 1 &amp; 0 \end{bmatrix}}_{\text{Transformation matrix in our language}} \begin{bmatrix} 2 &amp; -1 \\ 1 &amp; 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix}</math> </div>
<p>A transformation is applied to a space and transforms it. There is a vector in that space which is transformed. We want to express the transformation to eigenbasis (see next chapter first)</p>		
Initial Vector multiplied by the transformation that transforms the basis to eigenbasis, results in	Transformation expressed in initial basis times the prior, gives the transformed vector expressed in eigenbasis.	Inverse change of basis to eigenbasis, times the prior gives the transformed vector expressed in initial basis.

initial vector expressed in eigenbasis.

It is convenient to do the three matrices multiplication first that results in a new matrix which is the transformation expressed in eigenbasis (a diagonal matrix). this matrix can be multiplied (transform) by any vector expressed in eigenbasis,

## Eigenvectors and eigenvalues

Eigenvectors of a transformation are the vectors that remain on their own span after a transformation. Notice that they can be extracted for square matrices only. For non-square matrices we extract singular vectors and singular values.



Rotate  $30^\circ$  around  $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$

$$\begin{bmatrix} \cos(\theta) \cos(\phi) & -\sin(\phi) & \cos(\theta) \sin(\phi) \\ \sin(\theta) \cos(\phi) & \cos(\theta) & \sin(\theta) \sin(\phi) \\ -\sin(\phi) & 0 & \cos(\phi) \end{bmatrix}$$

Suppose a transformation is applied to a 2d space. All vectors are knocked off the lines that they span. All except some special vectors that remain on their own span. These are the eigenvectors of this transformation. All the vectors that lie in that line can be expressed as a linear combination of a unit vector on the line. Each eigenvector has a special value attached to it, the factor by which is stretched or squished after the transformation. This value is the eigenvalue of the eigenvector.

Why are eigenvectors important? In a 3d transformation, for example a 3d rotation, if you can find an eigenvector then you have found the axis of rotation. In this case the eigenvalue of that eigenvector would be 1 since a 3d rotation is a rigid body rotation which means that all the vectors of the space do not stretch or squish but they just rotate as they are.

So in many cases it is more convenient to think of a transformation in terms of eigenvectors and eigenvalues instead of the transformation matrix which gives too much weight to the coordinate system. The “eigen way” of description is less dependent on the coordinate system.

Matrix-vector multiplication

$$\overbrace{A\vec{v}}^{\text{Scalar multiplication}} = \underbrace{\lambda\vec{v}}$$

Scalar multiplication

This is the formula that describes an eigenvector  $v$ . It means that if you apply the transformation  $A$  to the space of  $v$ , the vector  $v$  will be just scaled by a factor (the eigenvalue). you have to solve for  $\lambda$  and  $v$ .

$$A\vec{v} = \lambda\vec{v}$$

$$A\vec{v} - \lambda I\vec{v} = 0$$

$$(A - \lambda I)\vec{v} = 0$$

$$\det(A - \lambda I) = 0$$

And this is the reasoning behind finding the eigenvectors and eigenvalues of a transformation.

First we have to represent  $\lambda$  vector multiplication with a matrix vector multiplication, where the matrix is such that when a vector is multiplied by it, will have the same effect as just scaling the vector. This matrix is a diagonal matrix which has  $\lambda$  in the diagonal and 0 everywhere else. This matrix is equal to  $\lambda$  multiplied by the Identity matrix (all zeros but ones in the diagonal). Then by doing algebra we end up in a new matrix ( $A-\lambda I$ ) that when multiplied by the vector  $v$  gives the zero vector. This equation is satisfied if the vector  $v$  is the zero vector but we want a non zero vector  $v$ . This means that the transformation associated with the  $A-\lambda I$  matrix transforms  $v$  into the zero vector. Since in a full rank transformation only the zero vector lands on the origin, this can only be true if this is not a full rank transformation. This means that it squishes space into a lower dimension which means that the unit object's volume of the initial space becomes zero, which means that the determinant of that matrix is zero. From the determinant equation we can find the eigenvalue  $\lambda$ . Since you know  $\lambda$ , you can solve the linear system of equations  $(A-\lambda I)v=0$  to find the eigenvectors  $v$ , which is the kernel (the null space) of the  $A-\lambda I$  matrix.

$$\det \begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix} = (3 - \lambda)(2 - \lambda) = 0$$

$$\lambda = 2 \text{ or } \lambda = 3$$

Seeking eigenvalue  $\lambda$

An example of a transformation.

For  $\lambda=2$  the solution to the linear system gives  $x=-y$  which means that the eigenvectors are all the vectors  $[y \ -y]$  which is the span of the vector  $[1 \ -1]$

For  $\lambda=3$  the solution to the linear system gives  $y=0$ . This means that the eigenvectors are all the vectors that have no  $y$ , which is the span of the vector  $[1 \ 0]$

$$\det \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = (-\lambda)(-\lambda) - (-1)(1) = \lambda^2 + 1 = 0$$

$$\lambda = i \text{ or } \lambda = -i$$

Interestingly, though, the fact that multiplication by  $i$  in the complex plane looks like a 90 degree rotation is related to the fact that  $i$  is an eigenvalue of this transformation of 2d real vectors.

A 2d transformation can have no eigenvectors. For example a 2d rotation by 90 degrees.

The eigenvalues for such a matrix are imaginary numbers. The fact that there are no real eigenvalues indicates that there are no eigenvectors.

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

A shear transformation

$$\det \begin{pmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{pmatrix} = (1-\lambda)(1-\lambda) = 0$$

$\lambda = 1$

Notice that a single eigenvalue can have more than a line full of eigenvectors. For example a scaling by a factor, here 2.

There is only one eigenvalue 2 but all the vectors of the space are eigenvectors since they remain in their own span. Solving the determinant gives  $\lambda=2$ . The linear system then becomes the zero matrix times v equals 0. Since any non zero vector multiplied by the zero matrix gives zero, eigenvectors are all the vectors of the space. If you want to select two eigenvectors that span the whole space its convenient to select [1 0] and [0 1]

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

## Eigenbasis

When both basis vectors are eigenvectors

$$\begin{bmatrix} -5 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

In a diagonal matrix, all the basis vectors are eigenvectors since if we think of the transformation that it describes, all basis vectors land on a scaled position within their span. They are just scaled. And the diagonal values are the eigenvalues since they are the factor by which they scaled.

$$\begin{bmatrix} 3^{100} & 0 \\ 0 & 2^{100} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Diagonal matrices are very convenient. You can very easily calculate the powers of a diagonal matrix. Its just a diagonal matrix with the eigenvalues (the diagonal values) raised to the power. If instead try to calculate the power of a non diagonal matrix, is a nightmare.

If a transformation happens to have enough eigenvectors so that a combination of some of them span the entire space, you can choose to use these eigenvectors as the basis vectors of the initial space. You just must make a change of basis, to express the transformation in relation with the new basis. Doing so the **transformation matrix** expressed in eigenbasis is guaranteed to be **diagonal** with the corresponding eigenvalues in the diagonal. It is so, since each basis vector lands on a scaled position within its initial span, so all the other coordinates of the transformed basis vectors are zero. The eigenbasis vectors are just scaled by the transformation.

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

Change of basis matrix

Use eigenvectors as basis

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$V^{-1}MV = \Lambda$$

M: a transformation of a space expressed in some basis of the space

V: a matrix whose columns are vectors that remain in their span after the transformation M is applied to the space (its columns are eigenvectors of M)

$\Lambda: I * \lambda$  (the eigenvalue matrix, a diagonal matrix with eigenvalues in the diagonal)

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

By  $V^{-1}MV$  we express the transformation M in relation to the new basis vectors which are eigenvectors of M. This will produce a new matrix, the matrix  $\Lambda$ , describing the same transformation but now the new matrix is diagonal.

Actually this way, we **diagonalizing matrix M**. the transformation defined by M is the same with the one defined by  $\Lambda$ . The only difference is that transformation  $\Lambda$  is expressed in an eigenbasis of M.

or  $M = V\Lambda V^{-1}$

M can be broken down into its eigenvector matrix times its eigenvalue matrix times the inverse of its eigenvector matrix.

So if you want to compute a power of a non diagonal matrix, it is much easier to transform the matrix to an eigenbasis (if it has enough eigenvectors), compute the power of the new matrix which is diagonal and then transform back to the initial basis. To do so we have to find the eigenvectors and eigenvalues of that transformation (we have to find the matrix V and then invert it). So we get the  $V^{-1}MV = M'$  where  $M'$  is diagonal.

Then we can process  $M'$  for example take the nth power  $M'^n$  and then transform back to the initial basis

$$VM'^nV^{-1} = M^n$$

### Eigenvectors for symmetric matrices

Symmetric matrices (a square matrix that is equal to its transpose, its rows are the same as its columns, whatever exist in one part of the diagonal exist also in the other) have orthogonal eigenvectors. As any matrix, a symmetric matrix S can be decomposed to a product of its eigenvector and eigenvalue matrices but in this case the eigenvector matrix and its inverse or transpose are orthogonal matrices. And therefore symmetric matrices are excellent for computations. Because they can be decomposed to orthogonal and diagonal matrices which are excellent for computations.  $S = Q\Lambda Q^{-1} = Q\Lambda Q^T$

### Singular values and singular vectors

Like eigenvectors but for non-square matrices.

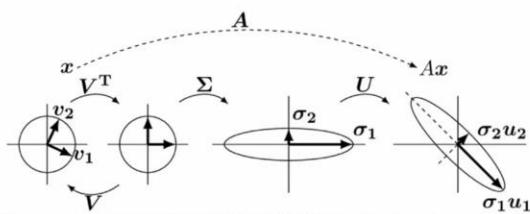
#### SINGULAR VALUE DECOMPOSITION

$$A = U\Sigma V^T \text{ with } U^T U = I \text{ and } V^T V = I$$

$AV = U\Sigma$  means

$$A \begin{bmatrix} v_1 & \dots & v_r \end{bmatrix} = \begin{bmatrix} u_1 & \dots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \text{ and } Av_i = \sigma_i u_i$$

SINGULAR VALUES  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$        $r = \text{rank of } A$



$U$  and  $V$  are rotations and possible reflections.  $\Sigma$  stretches circle to ellipse.

SVG is the best factorization of all, because it decomposes a matrix to its most important components by (and this is the important thing) decomposing it to two orthogonal and one diagonal matrices (these are the best matrices for computations)

U left singular vectors

V right singular vectors

$\Sigma$  singular values matrix

When the matrix A is square the U and V are identical and is the one eigenvector matrix.

We can make these singular vectors perpendicular to each other (linearly independent vectors) and that makes  $V^T$  and  $U$  orthogonal matrices.  $\Sigma$  is diagonal. So SVG combines the best of matrices.

Rotation ( $V^T$ ) + stretch ( $\Sigma$ ) + rotation ( $U$ )

That's what singular vectors and values do to a space.

The singular values are sequenced from largest to smallest. So  $\sigma_1$  is larger than  $\sigma_2$ .

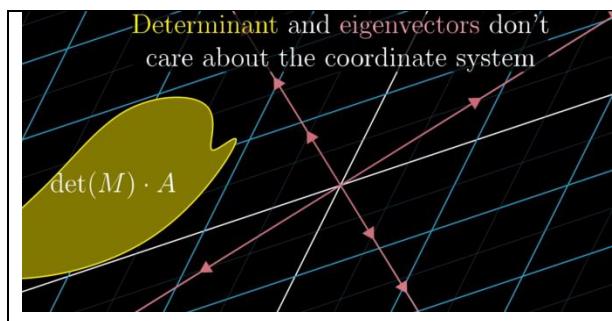
## Approximating a huge matrix using SVG

Using SVG you can get the important part of a matrix. Assume you have a huge matrix which is impossible to work with. You randomize the huge matrix and get a smaller sample from it (so it is more efficient for calculations). The random sampling is a good representation of the huge matrix. Then you apply SVG and you keep the larger singular values. So you can simplify your problem. The largest values are the important ones because these define the dominant part of the transformation (the largest stretching). The others are insignificant in comparison. So you can approximate the initial huge matrix with a simpler one (actually with three simple matrices  $U$ ,  $V$  and  $\Sigma$ ) by getting the most important features of it. Then you can make calculations extremely more efficient.

## Abstract vector spaces

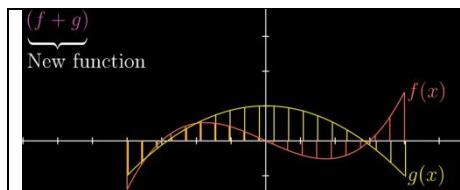
What are vectors at a fundamental level? List of Real numbers, vectors in a space or something else?

As you become more fluent with change of basis you start to realize that you are dealing with a space that exists, independently of the coordinates that you give it.



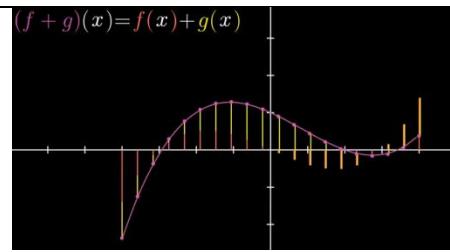
Thinks like **the determinant and eigenvectors are independent from the coordinate system**. They are inherently spatial. You can freely change the coordinate system without changing the underlying values for either one. But what is space really?

## Functions as vectors

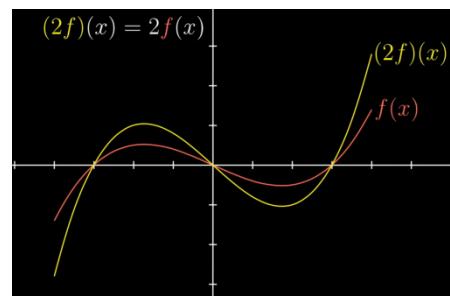


Functions can be added together and can be scaled. To add them you have to calculate the sum of the output value of each function in a specific location and do this for all infinite locations (inputs). To scale them you have to get the product of each output value with the scalar.

Given that the only thing vectors can do is to be added together and be scaled it seems that we can get the same problem solving techniques that we used in linear algebra

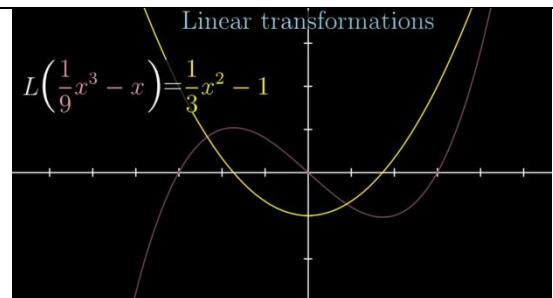


(linear transformations, null space, dot products, eigen everything etc.) and apply them to functions (which are actually vectors with infinite coordinates or rather vectors are functions with a specific set of input - output values).



If you think of the analogy, you can think of the coordinates of a vector as the output values of a function. The input value can be considered to be the position of the coordinate. You can think of a vector [7 0 5] as a function  $f$  that gets 3 inputs and gives 3 outputs.  $f(1)=7, f(2)=0, f(3)=5$ . You can think of a vector as equivalent to a specific function  $f$  (although  $f$  in reality would have values for each position in between). So a 3d vector is equivalent to a function with some specific properties. In this analogy, the whole 3d space of infinite 3d vectors is analogous with infinite functions. Another 3d vector would be equivalent to another function.

### Linear transformation of a function

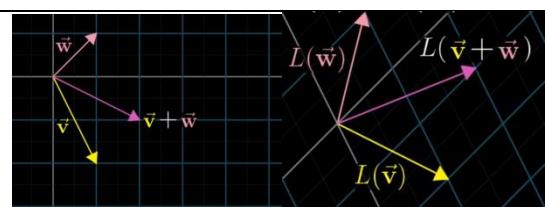


For example there is reasonable notion of a linear transformation of a function. But what does it mean for a transformation of functions to be linear? In general linearity has a somewhat abstract and symbolically driven definition. A transformation if linear it satisfies two properties, additivity and scaling (homogeneity):

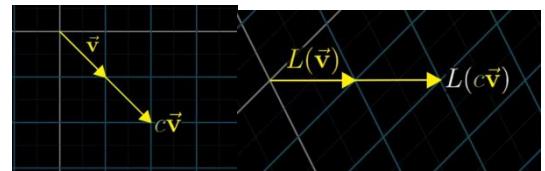
#### Formal definition of linearity

$$\text{Additivity: } L(\vec{v} + \vec{w}) = L(\vec{v}) + L(\vec{w})$$

$$\text{Scaling: } L(c\vec{v}) = cL(\vec{v})$$



**Additivity** means that if you add two vectors  $v$  and  $w$ , then apply a transformation on the sum you get the same result with adding the transformed versions of  $v$  and  $w$ .



**Scaling** means that if you scale a vector and then apply a transformation, you get the same result as if you scale the transformed vector by the same amount.

The way this is often described is that **linear transformations preserve addition and scalar multiplication**.

The illustration of grid lines that remain parallel and evenly spaced after a transformation while preserving the origin, is a consequence of the above statement, in the case which such a transformation is applied to vectors (points) of a 2d space.

An important consequence of a linear transformation is that it is completely described by where it moves the basis vectors. Since **any vector can be expressed by scaling and adding the basis vectors in some way, finding the transformed version of a**

**vector, comes down to scaling and adding the transformed versions of the basis vectors in that same way.** This is what makes matrix vector multiplication possible since it is translated numerically, to a matrix vector multiplication.

## Derivative operation as a linear transformation

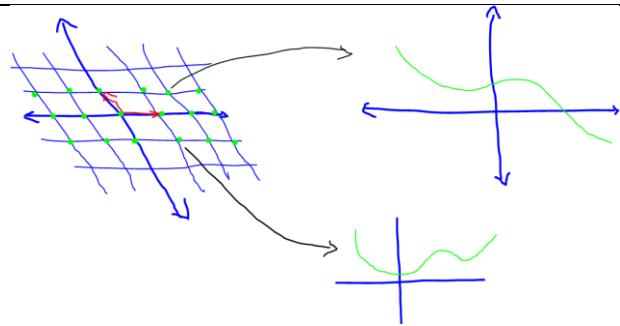
Derivative is linear

$$L(\vec{v} + \vec{w}) = L(\vec{v}) + L(\vec{w})$$

$$\frac{d}{dx}(x^3 + x^2) = \frac{d}{dx}(x^3) + \frac{d}{dx}(x^2)$$

$$L(c\vec{v}) = cL(\vec{v})$$

$$\frac{d}{dx}(4x^3) = 4\frac{d}{dx}(x^3)$$



The Derivative operation is a linear operation (operation, another word for describing a transformation) since it preserves both addition and scaling. Just as a 2d linear transformation does this for 2d vectors, the derivative operation (an infinite dimensions transformation) does for functions (vectors of infinite dimensions).

You can think of all functions (the derivative of which exists) as a space. A certain function would be a vector of that space. This space could be described by some basis vectors, or basis functions since a function is a vector. A transformation could be applied to that space, transforming it and all the vectors (functions) with it. If this transformation exhibits additivity and scaling then it is a linear transformation. The derivative operation is such a transformation and as such it can be represented by a matrix.

Our current space: All polynomials

$$\begin{aligned} & x^2 + 3x + 5 \\ & 4x^7 - 5x^2 \\ & x^{100} + 2x^{99} + 3x^{98} \\ & 3x - 7 \\ & x^{1,000,000,000} + 1 \\ & \vdots \end{aligned}$$

In this analogy each vector of that space (represented as a point where its tip is), is a distinct function. The basis vectors are functions themselves. All other functions can be expressed as addition of scaled basis functions. The derivation operation is a linear transformation of that space. A transformed function is the same linear combination of transformed basis functions. You can do all linear algebra stuff in this context. For example finding the eigenvectors or more appropriately, the eigenfunctions. An eigenfunction of the derivative operation is the function  $e^x$  since it remains the same after the transformation. Every function  $e^{kx}$  is an eigenfunction (where  $k$  is the eigenvalue as I think).

Let's try to create a matrix that describes the derivation operation. It seems tricky since function spaces tend to be infinite dimensional (if their domain is infinite). Let's define the space as the space of all polynomial functions. They have a finite number of terms but they can have infinitely large powers.

Next we have to define a basis. Since polynomials are written as addition of scaled powers of  $x$ , we can think of powers of  $x$  as the basis functions. The number of basis functions is infinite since we can have infinitely large powers of  $x$ , so every vector (polynomial) can be expressed as a list of infinite coordinates where 0 exist in all places of zero scaled powers.

Our current space: All polynomials

Basis functions

$$b_0(x) = 1$$

$$b_1(x) = x$$

$$b_2(x) = x^2$$

$$b_3(x) = x^3$$

$\vdots$

$$\underbrace{1x^2 + 3x + 5 \cdot 1}_{\text{Already written as a linear combination}}$$

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \\ a_n \\ 0 \\ \vdots \end{bmatrix} \quad \begin{array}{l} \text{Basis functions} \\ b_0(x) = 1 \\ b_1(x) = x \\ b_2(x) = x^2 \\ b_3(x) = x^3 \\ \vdots \end{array}$$

$$\frac{d}{dx}(1x^3 + 5x^2 + 4x + 5) = 3x^2 + \underbrace{10x + 4}_{\text{Basis functions}} \quad \begin{array}{l} b_0(x) = 1 \\ b_1(x) = x \\ b_2(x) = x^2 \\ b_3(x) = x^3 \\ \vdots \end{array}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 2 & 0 & \cdots \\ 0 & 0 & 0 & 3 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 5 \\ 4 \\ 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 4 \\ 2 \cdot 5 \\ 3 \cdot 1 \\ 0 \end{bmatrix}$$

$$1x^2 + 3x + 5 \cdot 1 = \begin{bmatrix} 5 \\ 3 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \quad 4x^7 - 5x^2 = \begin{bmatrix} 0 \\ 0 \\ -5 \\ 0 \\ 0 \\ 0 \\ 4 \\ \vdots \end{bmatrix}$$

In general terms since a polynomial has finite number of terms, its vector representation would be a finite list of numbers followed by an infinite tail of zeros. So if you want to construct a matrix that represents the derivation operation you can do it by taking the derivative of each basis function and putting the coordinates of the result in each column. This is the equivalent of where the basis vectors land after the transformation and is the matrix that describes the transformation from the view of the initial basis.

$$\frac{d}{dx} b_0(x) = \frac{d}{dx}(1) = 0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \frac{d}{dx} b_1(x) = \frac{d}{dx}(x) = 1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{d}{dx} b_2(x) = \frac{d}{dx}(x^2) = 2x = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$

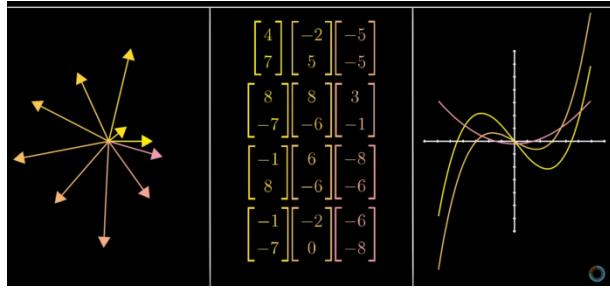
The result is a matrix with first column of 0s and a shifted diagonal of real numbers increased by one. If a polynomial is represented as a vector we can think that the derivation operation is applied to its space and transforms the vector to a new position. The resulting vector represents a polynomial too. This resulting polynomial is the derivative of the initial polynomial. If you do the multiplication you will see that this is indeed the derivative. This happens because the derivative operation is a linear transformation.

## Linear transformations

$$\begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Leftrightarrow \frac{df}{dx}$$

So taking the derivative of a function and matrix vector multiplication are members of the same family.

Linear algebra concepts	Alternate names when applied to functions
Linear transformations	Linear operators
Dot products	Inner products
Eigenvectors	Eigenfunctions



### Some comments

It's amazing how this makes everything add up. You can easily see how  $e^x$  is an eigenvector of the derivative transformation. While I was considering how you would go about calculating the determinant of an infinite matrix, I realized that it is just 0 because the first column is 0. Which makes sense since the derivative reduces a "dimension" from the polynomial (one of the transformed basis functions can be expressed as a linear combination of the other transformed basis functions). Which also explains why derivation is not invertible. For example the derivative of  $2x$  is 2. but you can't go back from 2 to  $2x$  because the derivative of  $2x+c$  is also 2. There is an infinite number of functions that have derivative 2.

I want to point out that there's actually a lot of depth to defining the invariants, i.e. geometric properties like the trace, determinant, set of eigenvalues, etc. of a linear transformation when you have infinitely many dimensions. For instance, with the determinant you have to multiply an infinite collection of numbers and you have to ask questions like: "When will this infinite product converge?" In finite dimensions you have a discrete set of eigenvalues, but for infinite dimensional transformations you can have a full continuum of eigenvalues as well. This deep interplay between **linear algebra** and **real analysis** is the subject of **functional analysis**.

## The 8 axioms of linear algebra

Rules for vectors addition and scaling

1.  $\vec{u} + (\vec{v} + \vec{w}) = (\vec{u} + \vec{v}) + \vec{w}$
2.  $\vec{v} + \vec{w} = \vec{w} + \vec{v}$
3. There is a vector  $\mathbf{0}$  such that  $\mathbf{0} + \vec{v} = \vec{v}$  for all  $\vec{v}$
4. For every vector  $\vec{v}$  there is a vector  $-\vec{v}$  so that  $\vec{v} + (-\vec{v}) = \mathbf{0}$
5.  $a(b\vec{v}) = (ab)\vec{v}$
6.  $1\vec{v} = \vec{v}$
7.  $a(\vec{v} + \vec{w}) = a\vec{v} + a\vec{w}$
8.  $(a + b)\vec{v} = a\vec{v} + b\vec{v}$

“Axioms”

In modern theory of linear algebra there are 8 axioms that a vector space must satisfy in order for all the linear algebra operations to be applied to it. They are just a check list that you can easily check to determine if the two fundamental properties of addition and scaling are preserved after a transformation.

So if you define some bizarre vector space that exhibits addition and scaling, if it satisfies these axioms then you can use the linear algebra tools to that vector space.

This abstract definition of vectors is the reason that every math book describes a linear transformation using addition and scaling in the definition rather than grid lines that remain parallel and evenly spaced. They don't want to be limited to a specific example. Abstractness is the price of generality.

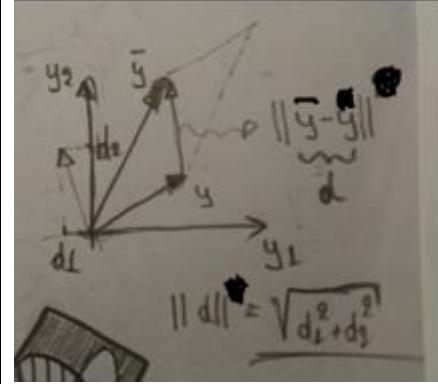
## Todo

- Transpose matrix
- Adjugate matrix
- Adjoint matrix
- Exponential matrix
- Trace
- Gaussian elimination and row echelon  
(it is an algorithm for solving systems of linear equations. So a method for finding the inverse of a matrix right?)
- Diagonalization
- Characteristic polynomial
- Minors and cofactors
- Invariants

## Misc

### Norm

In mathematics, a norm is a function from a real or complex vector space to the nonnegative real numbers that behaves in certain ways



## L2 norm (Second norm)

In particular, the Euclidean distance of a vector from the origin is a norm, called the Euclidean norm, or 2-norm, which may also be defined as the square root of the inner product of a vector with itself (which ultimately gives the Euclidean distance).

$$\|y\| = \sqrt{y^*y} = \sqrt{y_1^2 + y_2^2}$$

```

1 # l2 norm of a vector
2 from numpy import array
3 from numpy.linalg import norm
4 a = array([1, 2, 3])
5 print(a)
6 l2 = norm(a)
7 print(l2)

```

## Invariants

a function, quantity, or property that remains unchanged when a specified transformation is applied.

## Linear combinations

We use the term linear combination to describe any expression constructed from a set of variables by multiplying each variable by a constant and adding the results. (me: Since  $y=ax$  is the function of a line, the linear combination can be thought of as a combination of lines.)

## Formal linearity properties

$$L(\vec{v} + \vec{w}) = L(\vec{v}) + L(\vec{w})$$

$$L(c\vec{v}) = cL(\vec{v})$$

the transformation (function) L must exhibit these two properties in order to be considered linear.

A function f is linear if it obeys the following equation:

- (Additivity)  $f(ax_1 + bx_2) = af(x_1) + bf(x_2)$ .

If the input to the linear function f, consists of five parts  $x_1$  and three parts  $x_2$ , then the output of the function will consist of 5 parts  $f(x_1)$  and three parts  $f(x_2)$ . Essentially, linear functions transform linear combinations of inputs into the same linear combinations of outputs. *A function f is linear if it transforms linear combinations of inputs, into the same linear combinations of outputs.*

- (Homogeneity)  $f(\alpha \cdot x) = \alpha \cdot f(x)$

The general equation of a line (in 2 dimensions) is  $ax+by=c$  and even in multiple dimensions you can parameterize a line by a series of expressions like  $a_0x_0+b_0t=c_0, a_1x_1+b_1t=c_1, \dots, a_nx_n+b_nt=c_n$ . As such, terms with only one power of a variable (e.g.  $ax$  or  $by$ , but not  $xy$ ) are called "linear terms", since they are the terms you find in equations for lines.

Can we say that a transformation is linear if and only if every dimension is transformed linearly? For example in a 2 dimensions transformation (plane to plane for example tablet to wall), x should be transformed by a linear  $f(x)$  and y should be transformed by a linear  $g(y)$ . If so, then the transformation is linear and can be represented by a matrix. If it wasn't linear transformation (for example dimensions were transformed with  $x^2$ ) then you couldn't represent this transformation with a matrix. You should have a matrix for each point since each point is transformed differently

A linear function can be considered as a transformation function that maps two spaces with each other.

**Knowing the outputs of a linear transformation T for all “directions” (dimensions) in its input space is a complete characterization of T.** Without this linear structure, characterizing unknown input-output systems is a much harder task. Linear algebra is the study of linear structure, in all its details.

For example if T is a linear transformation that maps a two dimensional space to another two dimensional space (for example tablet to wall), then by mapping two lines of the tablet surface to the wall surface, one for each direction, we can completely describe T. first we create a point in the input space on  $(0,0)$  and see where is the  $(0', 0')$  in the output space. Then we map the  $(1,0)$  and the  $(0,1)$  which are the two directions. They are mapped to  $(1',0')$  and  $(0',1')$ . knowing these two we have a complete characterization of T. If we know that T is a linear transformation (by testing input output pairs for example -you need to have linearity in all dimensions, here two, as I understand) then any input to the linear function T can be described as a combination of  $(1',0')$  and  $(0',1')$ . For example here we want to find the mapping (projection) of the  $(2,3)$  vector of the input space:  $T(a, b) = aT(1,0) + bT(0,1)$

$$T(2,3) = T(2(1,0) + 3(0,1)) = 2T(1,0) + 3T(0,1)$$

### A Linear transformation can be represented by a matrix

y is a linear combination of the columns of A

$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \equiv \underbrace{\begin{bmatrix} x_1 a_{11} + x_2 a_{12} \\ x_1 a_{21} + x_2 a_{22} \\ x_1 a_{31} + x_2 a_{32} \end{bmatrix}}_{\text{row picture}} \equiv x_1 \underbrace{\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}}_{\text{column picture}} + x_2 \underbrace{\begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}}_{\text{column picture}}$	<p>y is a linear combination of the columns of A. This means that you can express any linear combination of a set of vectors as a matrix vector product.</p>
--	--

This is a core idea of linear algebra. Multiplication by a matrix A  $m \times n$  can be thought of as a linear transformation TA that takes n-sized vectors as inputs and produces m-sized vectors as outputs. In the above example your input is the x vector (size 2) which is transformed to a size 3 vector. Instead of writing  $y=TA(x)$  to denote the linear transformation applied to the vector x we write  $y=Ax$ . We say TA is represented by the matrix A.

The action of a function on a number is similar to the action of a linear transformation on a vector

<p>function <math>f : \mathbb{R} \rightarrow \mathbb{R} \Leftrightarrow</math> linear transformation <math>T : \mathbb{R}^n \rightarrow \mathbb{R}^m</math></p> <p>input <math>x \in \mathbb{R} \Leftrightarrow</math> input <math>\vec{x} \in \mathbb{R}^n</math></p> <p>output <math>f(x) \in \mathbb{R} \Leftrightarrow</math> output <math>T(\vec{x}) \in \mathbb{R}^m</math></p> <p>inverse function <math>f^{-1} \Leftrightarrow</math> inverse transformation <math>T^{-1}</math></p> <p>zeros of <math>f \Leftrightarrow</math> kernel of <math>T</math></p>	<p>function <math>f : \mathbb{R} \rightarrow \mathbb{R} \Leftrightarrow</math> linear transformation <math>T_A : \mathbb{R}^n \rightarrow \mathbb{R}^m</math> represented by the matrix <math>A \in \mathbb{R}^{m \times n}</math></p> <p>input <math>x \in \mathbb{R} \Leftrightarrow</math> input <math>\vec{x} \in \mathbb{R}^n</math></p> <p>output <math>f(x) \in \mathbb{R} \Leftrightarrow</math> output <math>T_A(\vec{x}) \equiv A\vec{x} \in \mathbb{R}^m</math></p> <p><math>g \circ f(x) = g(f(x)) \Leftrightarrow T_B(T_A(\vec{x})) \equiv BA\vec{x}</math></p> <p>function inverse <math>f^{-1} \Leftrightarrow</math> matrix inverse <math>A^{-1}</math></p> <p>roots of <math>f \Leftrightarrow</math> kernel of <math>T_A \equiv</math> null space of <math>A \equiv \mathcal{N}(A)</math></p> <p>image of <math>f \Leftrightarrow</math> image of <math>T_A \equiv</math> column space of <math>A \equiv \mathcal{C}(A)</math></p>
<p><b>Table 2.1:</b> Correspondences between functions and linear transformations.</p>	

Abstract vectors: Things that can be added. As I understand it, vectors can be thought of as operations to a state, that when applied together to this state have the same effect with when they are applied to it one after the other? For example, a wave can be represented like a list of three numbers (wavelength, magnitude, phase) and two waves can be added together. (so as I understand, you can represent a wave with a 3d vector and make calculations between vectors to find the interaction of waves.)

## Vector operations

addition, subtraction, scaling, dot product and cross product

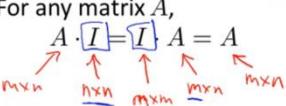
## Matrix operations

addition, subtraction, scaling, matrix product ( $AB$ ), matrix vector product ( $Av$ ), matrix inverse ( $A^{-1}$ ), trace ( $\text{Tr}(A)$ , determinant ( $\det(A)$  or  $|A|$ )

Inverse:  $A^{-1}(A\vec{x}) = A^{-1}A\vec{x} = \vec{x}$ .

The cumulative effect of applying  $A^{-1}$  after  $A$  is the identity matrix (μοναδιαίος πίνακας):

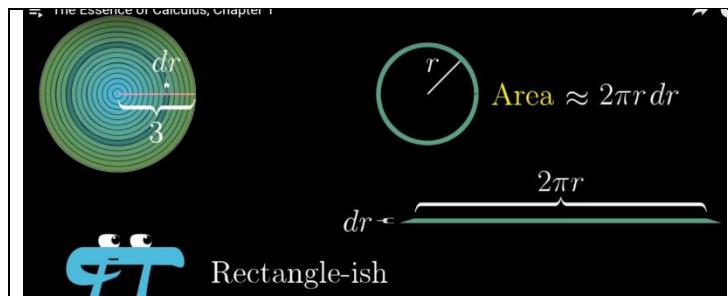
$$A^{-1}A\vec{x} = \mathbb{1}\vec{x} = \vec{x} \Rightarrow A^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1.$$

<p>For any matrix <math>A</math>,</p>  $A \cdot \boxed{I} = \boxed{I} \cdot A = A$	<p>Identity matrix</p> <p><math>A^*I=I^*A=A</math> where <math>I</math> is the identity matrix</p>
<p><b>Matrix inverse:</b> If <math>A</math> is an <math>m \times n</math> matrix, and if it has an inverse, <math>AA^{-1} = A^{-1}A = I</math>.</p>	<p>Inverse matrix Have in mind that not all numbers have an inverse. 0 doesn't have. Only square matrices can have an inverse but not all have. Those that don't are called singular or degenerate.</p>
<p><b>Matrix Transpose</b> Example: <math>A = \begin{bmatrix} 1 &amp; 2 &amp; 0 \\ 3 &amp; 5 &amp; 9 \end{bmatrix}</math>      <math>A^T = \begin{bmatrix} 1 &amp; 3 \\ 2 &amp; 5 \\ 0 &amp; 9 \end{bmatrix}</math> Let <math>A</math> be an <math>m \times n</math> matrix, and let <math>B = A^T</math>. Then <math>B</math> is an <math>n \times m</math> matrix, and <math>B_{ij} = A_{ji}</math>.</p>	<p>Transpose matrix</p>

A 4 dimensional vector belongs to the set of  $R^4$ , the set of all 4d vectors with real numbers as values.

# The essence of Calculus

## Introduction



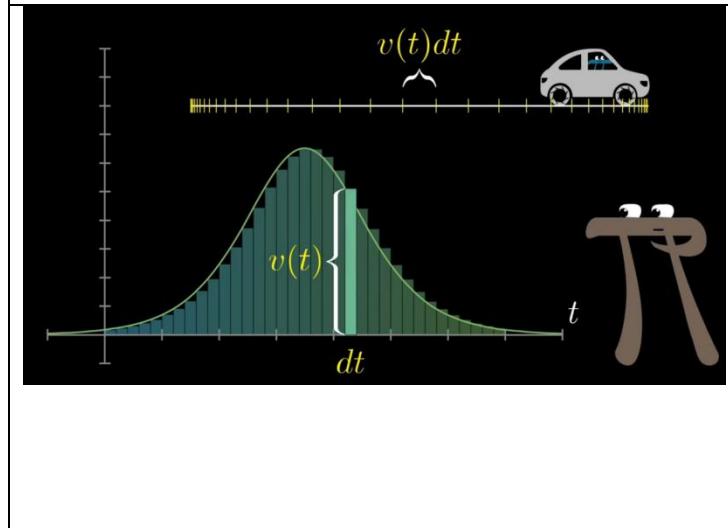
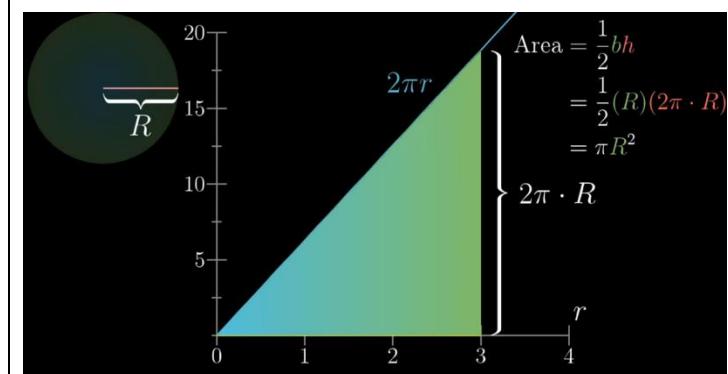
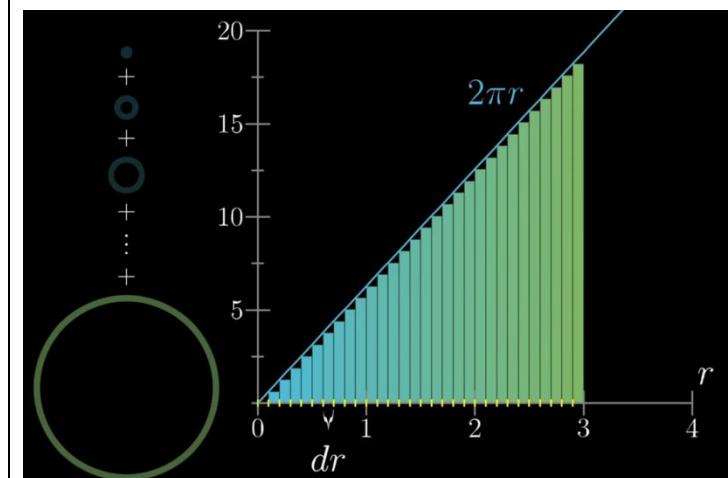
Imagine that we want to find a formula that gives the area of a circle. We can split it to different areas. One kind of splitting is to split the circle area to concentric circle. This splitting makes sense since it respects the symmetry of the circle and Math has a tendency to reward you when you respect its symmetries. You can try to approximate the area of such a slice considering it as rectangle-ish. The smaller the thickness, the more true this would be (the more rectangular the slice would be) since top and bottom sides would tend to be equals. Each slice has an area of  $2\pi r * dr$  and the area of the circle is the sum of all these areas.

We can create a cartesian plane with x axis the radius of a slice. Approximating a slice with a rectangular shape, we can place them all together on the x axis. The sum of their area would be the area of the circle.

The thickness of the slice is  $dr$  the height is  $2\pi r$ . The x axis is the  $r$  the y axis is the height which is a function of  $r$  so we can express it as a function  $f(r)=2\pi r$  and graph it. Doing so, we can clearly see that the sum of many small slices approximates the area under a graph. The smaller the  $dr$  the closer we would be to the area under the graph. So we can deduce that the exact (not an approximate) answer is the area under the graph. **We can deduce that the value that our formula (sum of many small values) approaches without never reaching, is the exact answer.** Since the graph forms a triangle its area equals to  $\text{Area} = \frac{1}{2} \cdot b \cdot h$  which is  $\frac{1}{2} \cdot 3 \cdot 2\pi^3$  or  $\pi r^2$  in general.

Many problems can be approached like this, as a sum of small values. It turns out that most of these problems are equal to the area under a graph. **This happens whenever the quantities that you add, the sum of which approximates the quantity that you are looking for, can be thought of as the areas of many thin rectangles sitting side by side.**

In order for a quantity to be able to be represented by the area of a rectangle  $b*h$ , it needs to be able to be expressed as a product of two quantities where one quantity is the one across which we slice and the quantity that we are looking for is a function of the sliced quantity.



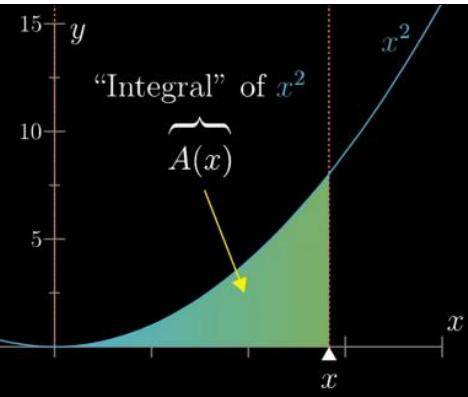
For example finding the distance traveled by an object if you know its velocity at each point in time.

## Integral and derivative

The Integral of a function F(x) at a certain input x, represents the value to which a certain sum converges. The value to which the sum of the “signed” areas of a large number of small parallelograms up to a point x under the function graph converges to, as the side of the parallelograms becomes smaller and smaller is called the integral of that function f at that specific input x. Since the integral value depends on the input x, the integral is a function of x too.

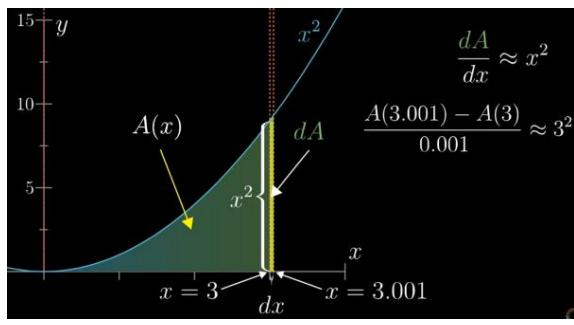
The derivative of a function A(x) at a certain input x, represents the value to which a certain ratio converges. We know the function F(x) but we don’t know the function that describes the area under its graph (its integral A(x)). What we do now, is how the area is affected if we make a small change to the input value x of the function f. This value can be represented by a ratio and is called the derivative of the function that describes the area (the derivative of the integral) at that specific point. Since the derivative value depends on the input x, the derivative is a function of x too.

The derivative of the integral of a function at a certain input is equal to the value of the function at that input. This is the fundamental theorem of calculus.



This technique is important since many problems that can be expressed as the sum of many small things can be reframed as finding the area under a graph.

The area under a graph between 0 and a random value  $x$ , is a function of  $x$  and is called integral of the graph’s function. (the value which the sum of the area of many small slices, approaches)



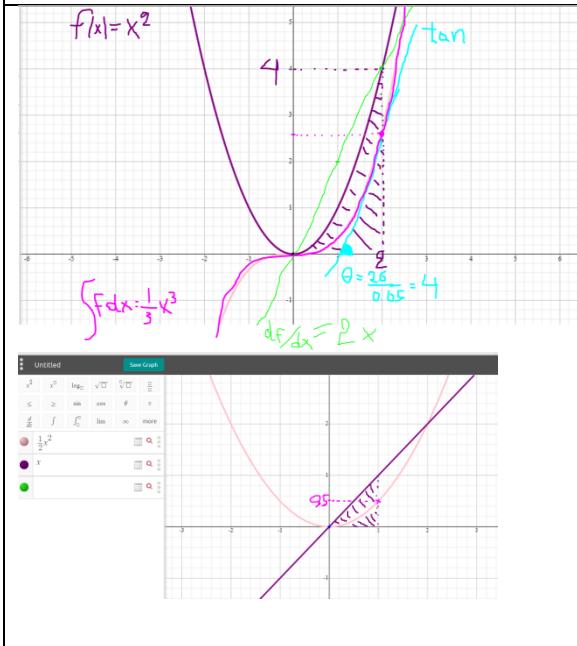
Finding the area under a graph is genuinely hard. For example finding the area under the graph of  $x^2$ , one way to approach this, is to think of what happens to  $A(x)$  if you slightly increase  $x$ . The smaller the  $dx$ , the better a rectangular area approximates the change in  $A$ . The ratio of a tiny change in  $A$  to the tiny change of  $x$  that caused it, is equal to whatever  $x$  was at that point, squared. We don’t know  $A(x)$  but we know one property that this function must have.

This applies to any random function  $f(x)$  the graph of which has an area  $A(x)$ . Giving a specific input  $x$  to the function  $A(x)$  results in a certain output value (a certain area). If you slightly change this input, the output changes slightly too. This tiny change to the output of  $A$  divided by the tiny change

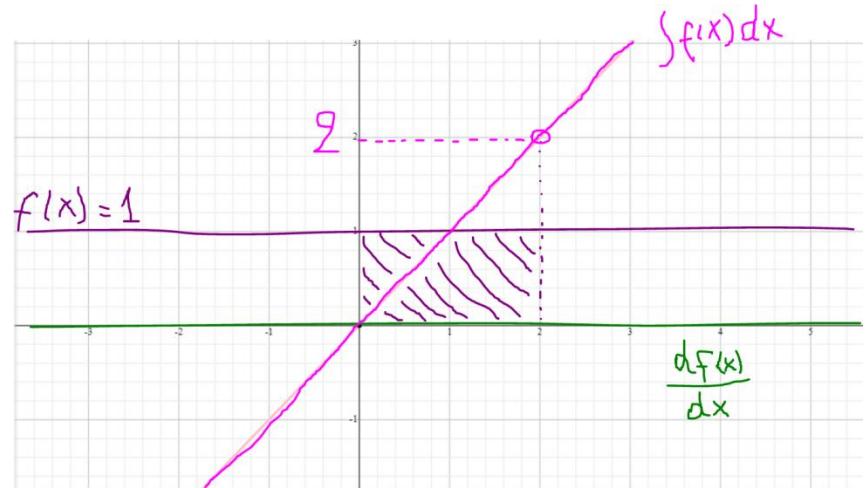
$$\frac{dA}{dx} \approx x^2$$

to the input that caused it, is about equal to the height of the graph that forms A, at the initial input, the output of  $f(x)$ . This approximation gets better as the change gets smaller.

This ratio shows how much the area will be affected by a small change in  $x$ . this is the definition of the derivative of the function that relates the area with  $x$ . More formally, **the derivative is whatever this ratio approaches as the slight change gets smaller and smaller. The derivative is a measure of how sensitive a function is to a change in its input.** In this case the function is  $A(x)$  which gives the area under the graph, We don't know the function  $A(x)$  but we know that its derivative is  $x^2$ .



Since the area that a function  $F(x)$  forms with x axis, is a function of  $x$  itself, it has its own graph too. The area under the  $F$  graph at a point  $x_1$ , is the value of the function  $A$  at  $x_1$ . the derivative of the function  $A$  is the original function  $F$ . The derivative of  $F$  is a different function.



## The fundamental theorem of calculus

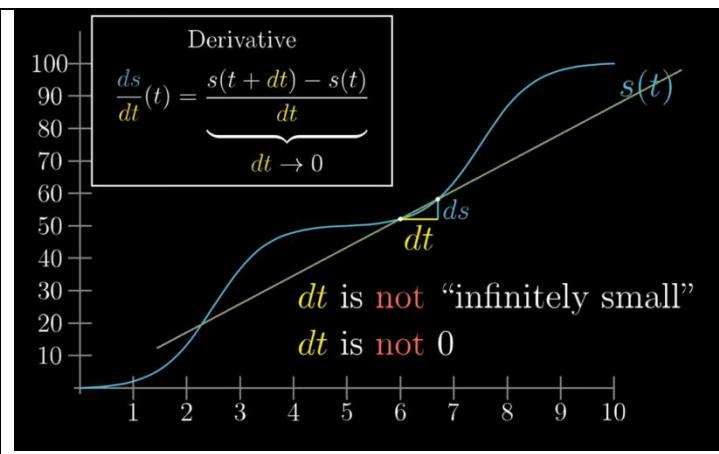
The derivative of a function that describes the area under a graph, gives you back the function that describes the graph itself. It ties together integrals and derivatives and shows how in one sense one is the inverse of the other (you integrate a function

## Derivative

The derivative of a function at a specific point, is the best constant approximation of the rate of change of the function around that point.

Derivatives is fundamentally a way to look at small changes in some quantity and how that relates to a resulting change in another quantity. The ratio of the resulting change to the change that caused it, is the rate of change of the resulting quantity (per unit change of the initial quantity)

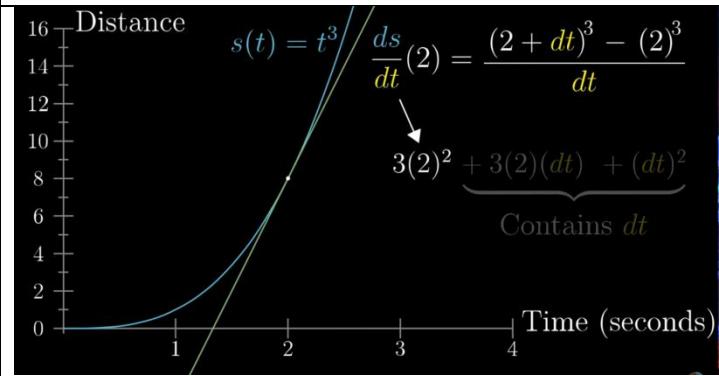
## The paradox of the derivative at specific points



The derivative is the value that the ratio approaches as  $dt$  goes to 0, it is not the a ratio itself, since if we wanted to get the exact value at a particular point,  $dt$  should have been zero and the ratio wouldn't be defined.

The pure math definition is that the derivative is the slope of the tangent to a graph at a specific point. The bigger the change ( $ds$ ) the bigger the slope of the tangent.

This way we can flirt with the paradox of change in an instance without ever touching it.



This is how the derivative formulas are calculated. They arise algebraically from the ratio  $df/dt$  at certain points. Since we want to get the value which this ratio approaches as  $dt$  becomes 0, all terms with  $dt$  can become 0. what lefts is the approached value.

$$\frac{ds}{dt}(2) = 3(2)^2$$

If  $dt$  has a specific size.

$$\frac{d(t^3)}{dt} = \frac{t^3 + 3t^2 dt + 3t(dt)^2 + (dt)^3 - t^3}{dt}$$

**Messy**

$dt \rightarrow 0$

$$\frac{d(t^3)}{dt} = 3t^2$$

**Simple**

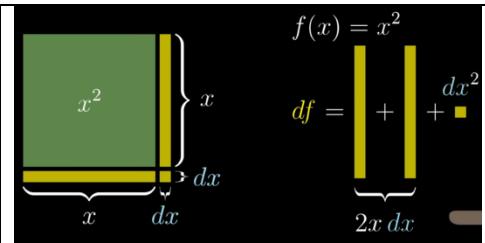
A good way of thinking about the derivative of a function in a certain point, is that **it is the best constant approximation of the rate of change of this function around that point**. At time  $t=0$ , is the car moving? The derivative of the distance to time function is speed to time, and the derivative formula gives speed=0 in  $t=0$ . So if it doesn't start moving at  $t=0$  then when does it start? This is a paradox and the proper answer is that the question makes no sense. It references the idea of change in a particular moment ( $t=0$ ) but there can't be no change in a particular moment. You need to compare it to another moment so that the notion of change has a meaning. What this means based on derivative definition, is that the best constant approximation of speed around 0 is 0.

## Derivative Rules

Power rule, add rule, product rule, chain rule

### Power Rule

You can visualize the various derivative formulas through geometry



Suppose that a rectangle represents the function  $f(x)=x^2$ . If we increase the side  $x$  by a small amount  $dx$  the area changes too, by a quantity represented as  $df$ .

$$\begin{aligned} \frac{d(x^1)}{dx} &= 1x^0 \\ \frac{d(x^2)}{dx} &= 2x^1 \\ \frac{d(x^3)}{dx} &= 3x^2 \\ \frac{d(x^4)}{dx} &= 4x^3 \\ \frac{d(x^5)}{dx} &= 5x^4 \end{aligned}$$

"Power rule"

$$\frac{d(x^n)}{dx} = nx^{n-1}$$

You can safely ignore any quantity that is  $dx$  raised to a power bigger than one since it would be really small. This can be algebraically proved since when we divide by  $dx$ , the  $dx^2$  turns to  $dx$  and is added to the result. But since  $dx$  approaches 0, the contribution of that term is zero.  $[df=2x*dx+dx^2 \rightarrow df/dx=2x+dx \rightarrow dx \rightarrow 0 \rightarrow df/dx=2x]$

So it is safe to say that  $df=2x*dx$  so  $df/dx=2x$  meaning that the rate of change of the area of the rectangle as the side changes is two times the side. If the side is 1 then a increase of the side by one unit of length, results in an increase in the area of 2 units of area.

Generalizing, the derivative of polynomial terms give the power rule.

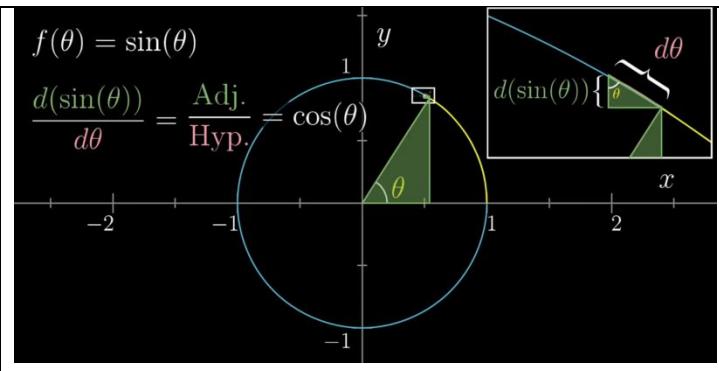
$$\begin{aligned} x^n &\downarrow \\ (x + dx)^n &= \underbrace{(x + dx)(x + dx)(x + dx) \cdots (x + dx)}_{n \text{ times}} \\ \text{Nudged output} &= x^n + \cancel{nx^{n-1}dx} + (\text{Multiple of } dx^2) \\ &\quad + dx \cdot x \cdot x \cdots x \\ &\quad + x \cdot dx \cdot x \cdots x \\ &\quad + x \cdot x \cdot dx \cdots x \\ &\quad \vdots \\ &\quad + x \cdot x \cdot x \cdots dx \end{aligned}$$

How the power rule is proved algebraically

Initially we have a quantity  $x^n$ . We apply a small change in  $x$ , so the resulting quantity would be  $(x+dx)^n$ . We want to find how the quantity changed. To do so we can get the difference between after and before the change. Doing algebra on the changed quantity, it turns out that all but a negligible portion of the increase in the output comes from the  $nx^{n-1}$  term.

So  $df=x^n-(x+dx)^n=nx^{n-1}dx+sth*dx^2 \rightarrow df/dx=nx^{n-1}+sth*dx$  As  $dx>0$   $sth*dx>0$  too.

That is why the derivative of  $x^n$  is  $nx^{n-1}$ .



$$d(\sin\theta)/dx = \cos\theta$$

$$d(\cos\theta)/dx = -\sin\theta$$

### Derivatives of combinations of functions

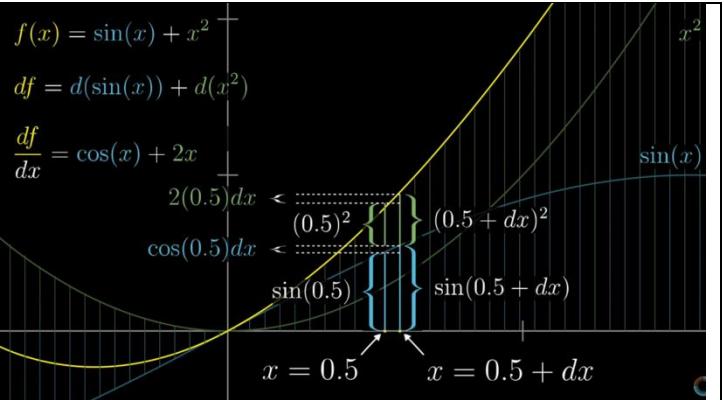
There are 3 basic ways to combine functions: add them together, multiply them and compose them (throw one inside the other). Subtracting or dividing can be expressed with the three basic ways.

## Sum Rule

Sum rule

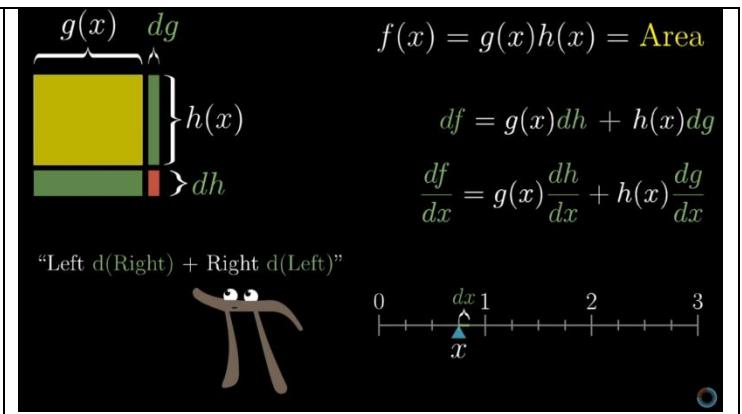
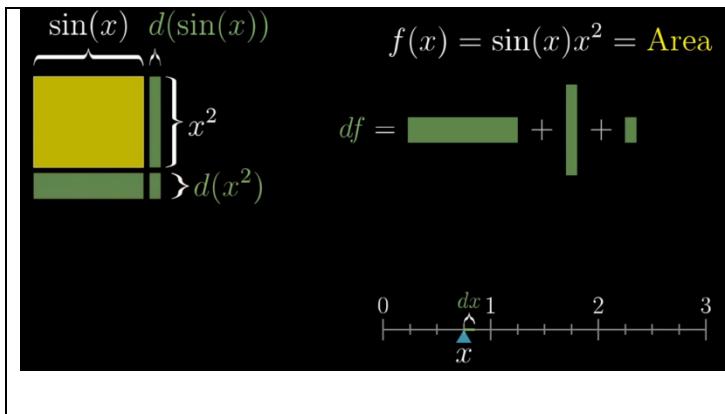
$$\frac{d}{dx}(\sin(x) + x^2) = \cos(x) + 2x$$

$$\frac{d}{dx}(g(x) + h(x)) = \frac{dg}{dx} + \frac{dh}{dx}$$



## Product Rule

Usually the best way to visualize a product of two things is as some kind of area. For example the area of a box in which each side's length is a function of x. As you change x, the sides length change accordingly. This way the product of the two functions can be represented by the area of that box. So to see the derivative, you can see how this area is affected by a tiny increase in x.



## Chain Rule

“Chain rule”

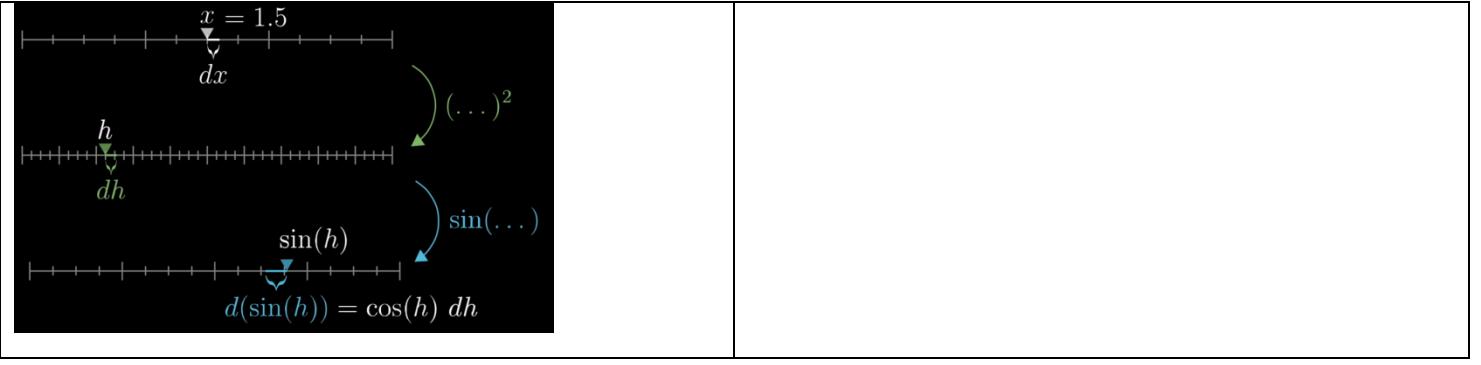
$$\frac{d}{dx} \sin(x^2) = \cos(x^2) 2x$$

$$\frac{d}{dx} \underbrace{g(h(x))}_{\text{Outer}} = \underbrace{\frac{dg}{dh}(h(x))}_{\text{Inner}} \underbrace{\frac{dh}{dx}(x)}_{d(\text{Outer})}$$

We express the change of the third quantity relative to the second one (dh). we have already expressed the second in relation to the first so at the end we can unfold everything.

$$dh = d(x^2) = 2x dx$$

The derivative of g evaluated on h, multiplied by the derivative of h.



## Derivatives of exponentials

All exponential functions are proportional to their own derivative.

$$M(t) = 2^t \quad M(t) = 2^t$$

$$\frac{dM}{dt}(t) = \frac{2^{t+dt} - 2^t}{dt} \quad \frac{dM}{dt}(t) = \underbrace{\frac{2^t 2^{dt} - 2^t}{dt}}$$

$$M(t) = 2^t$$

$$\frac{dM}{dt}(t) = 2^t \left( \frac{2^{dt} - 1}{dt} \right)$$

$$\frac{2^{0.001} - 1}{0.001} = 0.6933875\dots$$

We want to calculate the derivative of the exponential function with base two  $2^t$ . Writing the ratio and making some algebra we end up with a formula which shows that the derivative is a product of the exponential function itself with some quantity that is a function of  $dt$ .

$$\frac{2^{dt} - 1}{\underbrace{dt}_{dt \rightarrow 0}}$$

As  $dt$  approaches to zero, this quantity approaches a certain value which is different for each base. So  $d/dt(a^t) = c * a^t$  where  $c$  is a constant that depends on the base  $a$ .

It turns out that this constant equals to 1 if the base of the exponential is the number e. Actually this is another definition for e (the number to which that formula approaches when dt approaches to zero). We can find the derivative of any multiple power of e applying the chain rule.

$$\frac{de^{ct}}{dt} = \frac{de^{ct}}{dct} \cdot \frac{dct}{dt} = e^{ct} \cdot c$$

If we could express any exponential with base  $a$  as another exponential with base  $e$ , we could easily calculate its derivative.

Suppose we want to calculate the derivative of  $a^x$ . We can try to express  $a$  as a power of e. We want to find a number  $x$  so that  $e^x=a$ . Solving this equation for  $x$  ( $\ln x$  is the inverse function of  $e^x$ ) we get that  $x$  should be the natural logarithm of  $a$ .

$$e^x = a \Rightarrow \ln e^x = \ln a \Rightarrow x = \ln a$$

This means that

$$a = e^{\ln a}$$

So any exponential with base  $a$  can be rewritten as an exponential with base  $e$ . This is a very convenient fact. **When we work with the rate of change of exponentials, we should convert them to exponentials of base e.**

$$a^x = (e^{\ln a})^x = e^{\ln a \cdot x}$$

This means that the derivative of any exponential with base  $a$  would be

$$\frac{da^x}{dx} = \frac{de^{x \ln a}}{dx} = \frac{de^{x \ln a}}{dx \ln a} \cdot \frac{dx \ln a}{dx} = e^{\ln a \cdot x} \cdot \ln a$$

For the derivative of an exponential with base two  $2^x$ , this formula gives  $2^x = e^{\ln 2 \cdot x}$  so  $d(2^x)/dx = \ln 2 \cdot e^{\ln 2 \cdot x} = \ln 2 \cdot 2^x$

Which shows that the **proportionality constant** of the derivative is the natural logarithm of 2 ( $\ln 2 = 0.69315\dots$ ) and is the quantity which is a function of  $dt$  that we described before. Generalizing we can see that the derivative of an exponential with base  $a$  is the same exponential multiplied by the natural logarithm of its base.

$$\frac{d(a^x)}{dx} = \ln a \cdot a^x$$

$$2^t = e^{\overbrace{(0.69315\dots)t}^{\log_e(2)}} = \pi^{\overbrace{(0.60551\dots)t}^{\log_{\pi}(2)}} = 42^{\overbrace{(0.18545\dots)t}^{\log_{42}(2)}}$$

Notice that similarly we can express any exponential to an exponential with a base of our choice.

Apart from the mathematical convenience, writing exponentials in the form with base e, gives the constant in the exponent a nice meaning.

For example if we examine some phenomena like these:

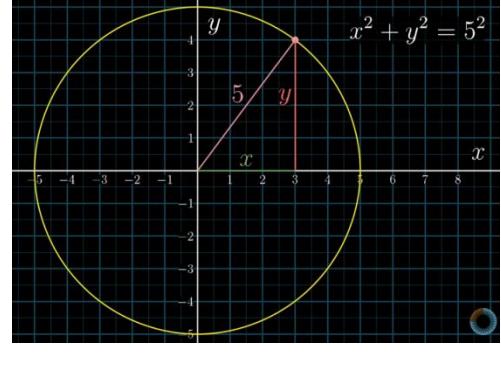
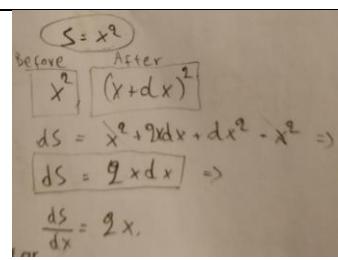
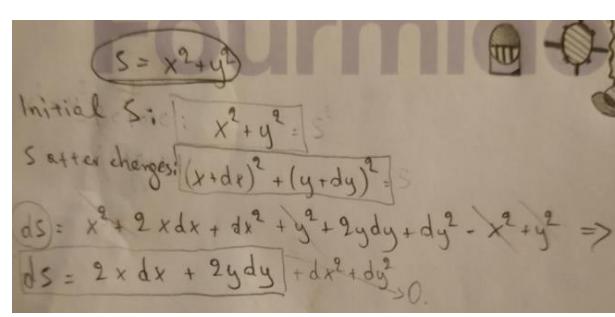
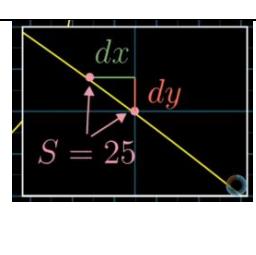
- The rate of change of a population is proportionate to itself (to the population at any given time)

- The rate of change of the temperature of a hot liquid in a room, is proportionate with the difference of temperatures between the room and the water. Or said a little differently, the rate by which this difference changes is proportionate to itself (to the difference at any given time).
- The rate at which an investment grows is proportionate to itself (to the money at any given time)

In all these cases the rate of change of a variable, is proportionate to itself (to the variable at any given time). All these phenomena (the functions that describe these variables in relation to time) can be described with exponential functions since the rate of change of an exponential function is proportionate to itself. And it is preferable to express these functions as exponentials with base  $e$  and exponent a constant times  $t$  since this constant at this form has a natural meaning. It is equal to the proportionality constant of the rate of change. A function  $a^t$  has a rate of change  $\ln a * a^t$ . But looking at the function in the form  $a^t$  we can't understand much about its rate of change. If we write it instead with base  $e$ , it would be  $e^{\ln a * t}$  and just by looking at it we can deduce the proportionality constant of its rate of change. It is the constant in its exponent ( $\ln a$ ).

## Implicit differentiation

Whenever you have a relationship between two variables, but in such a way that one isn't a function of the other, then you write the relationship as an equation that has both variables in the left side (the other side could have a variable too). There is no way to simplify it more or to write it in a  $y=f(x)$  form (since  $y$  is not a function of  $x$ ). The graphs of such equations are called implicit curves and the process of finding the derivative of such equations is called implicit differentiation in which the change of one variable is related to the change of the other (related rates)

	<p>The equation of a circle is not a function. Its graph is just an implicit curve. Implicit curve is the set of all points <math>(x,y)</math> that satisfy some property written in terms of the two variables. <math>x</math> is not an input, <math>y</math> is not an output. They are just interconnected values related with one equation. The implicit differentiation formula gives:</p> $x^2 + y^2 = 5^2$ $2x \, dx + 2y \, dy = 0$ $\frac{dy}{dx} = -\frac{x}{y}$
	<p>But where did it come from? We can name <math>S</math> the quantity described by the equation. We can think of the derivative of the quantity as the rate that this quantity changes from a small change in <math>x</math> and a small change in <math>y</math>. Doing algebra we end up with this formula.</p>
	<p>When we want to find the change of <math>S</math> where <math>S</math> is a function of <math>x</math>, then we end up in a changed quantity described by an equation like: <math>ds = f(x)dx</math> which means that we can get an expression for the rate of change of <math>S</math>: <math>ds/dx = f(x)</math>.</p>
	<p>When the quantity <math>S</math> is not a function of <math>x</math> but an expression of two independent variables <math>x</math> and <math>y</math>, the change of <math>S</math> is an expression of <math>dx</math> and <math>dy</math> and as such it can't be expressed as a ratio of <math>dS</math> over one change (there are two changes). And the two changes are related to one another. If we change the <math>x</math> by a small <math>dx</math>, we must change the <math>y</math> by a specific <math>dy</math> so that we are still in the graph of the equation. <math>dy</math> is related to <math>dx</math> (related rates). <b>Related rates</b> can be expressed as an equation if</p>

	<p>we solve for <math>dy/dx</math> (<math>-x/y</math> in the case of the circle). <b>It shows us how much <math>dy</math> must be in order for the <math>x</math> and <math>y</math> equation to be respected</b> (in case of circle, the resulting position after <math>dx</math> and <math>dy</math> to still be in the circle).</p>
	<p>In case of <math>S</math> (or <math>f</math>) as a function of <math>x</math> instead, you change the <math>x</math> by <math>dx</math> and you try to find what is the <math>df</math>, how that <math>dx</math> causes the function to change.</p>

## Related rates

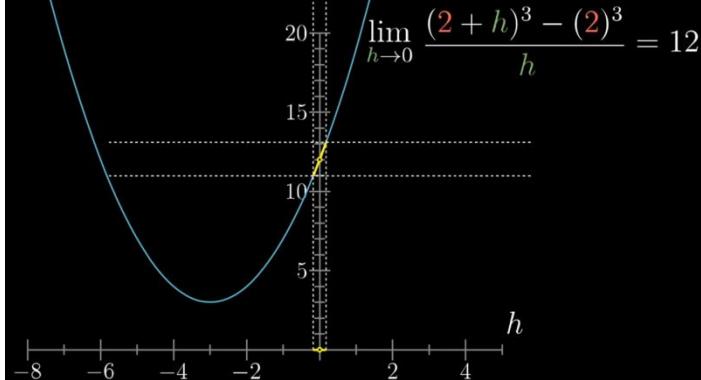
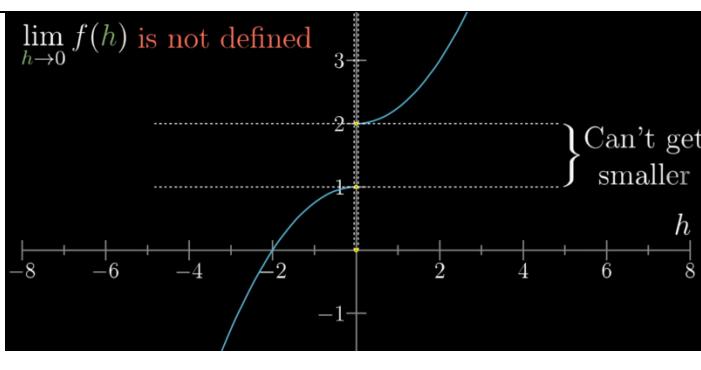
<p>Related rates</p> $x(t)^2 + y(t)^2 = 5^2$ $\frac{d(x(t)^2 + y(t)^2)}{dt} = 0$ $2x(t)\frac{dx}{dt} + 2y(t)\frac{dy}{dt} = 0$	<p>Related rates</p> $x(t)^2 + y(t)^2 = 5^2$ $x(t) = (5^2 - y(t)^2)^{1/2}$ <p>Find <math>\frac{dx}{dt}</math></p>
$2(3)\frac{dx}{dt} + 2(4)(-1) = 0$ $\frac{dx}{dt} = \frac{4}{3}$	<p>There is though another way to do the same thing. The left part of the equation is an expression which is a function of time (although constant). The derivative of this expression is the rate at which the expression changes as time changes. A small change in time <math>dt</math> causes a small change in <math>y</math> (<math>dy</math>) and a small in <math>x</math> (<math>dx</math>). So ultimately a small <math>dt</math> causes a small change to the expression.</p> <p>So we can take its derivative over time to see how it changes over time. Doing algebra we can find the <math>dx/dt</math> rate of change.</p>

<p><math>S(x, y)</math></p> $x^2 + y^2$ $dS = 2x dx + 2y dy$	<p>This looks like the derivative of the circle but in case of circle there is no notion of change in time. Just changes in <math>dx</math> and <math>dy</math>, not <math>dt</math>. A way to approach this is to think of the quantity <math>S</math> as a function of two variables.</p> <p>The key think is to restrict the small changes in such a way that the resulting position is still in the circle. This is what makes the differentiation implicit</p>
	<p>But when you restrict yourself in small changes that keep you on the circle, then the quantity <math>S</math> doesn't change and consequently its derivative <math>dS</math> would be 0.</p> <p>This condition keeps you in the tangent of the circle not the circle itself but for small enough steps the tangent line is identical to the circle.</p>

## Limits

They allow us to avoid talking about infinitely small changes by instead asking what happens as the size of change to a variable approaches zero.

<p>Formal derivative definition</p> $\frac{df}{dx}(2) = \lim_{h \rightarrow 0} \frac{f(2 + h) - f(2)}{h}$	<p>The <math>df/dx</math> ratio is almost what a derivative is. The actual derivative is whatever this ratio approaches when <math>dx</math> approaches 0.</p>
---	--

	<p>Graphing the limit as a function of <math>h</math></p> $f(h) = \frac{(2+h)^3 - 2^3}{h}$ <p>we can see that this function seems continuous but it is not. For <math>h=0</math> the output value is <math>0/0</math> which is not defined. This is represented in the graph with an empty point (a hole). We can clearly see though that as <math>h</math> approaches 0 no matter from what side it approaches it, the output value approaches 12.</p>
<p><math>\lim_{h \rightarrow 0} f(h)</math> is not defined</p> 	<p>In this case though, the limit is not defined. The limit approaches a range of length 1, not a specific value.</p>

$\epsilon, \delta$  definition of limits is a formal definition of what a limit is

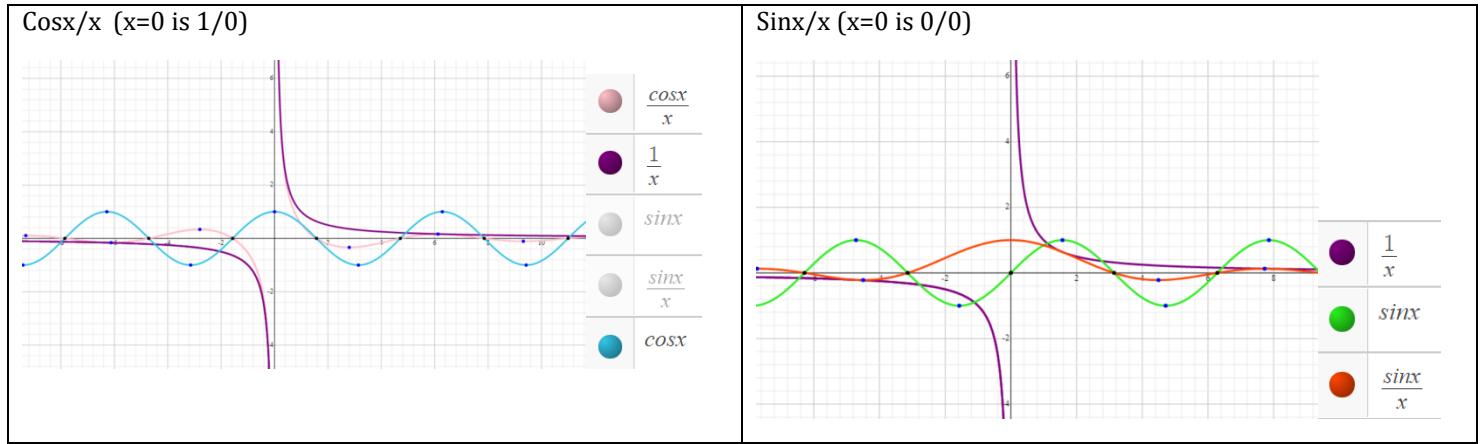
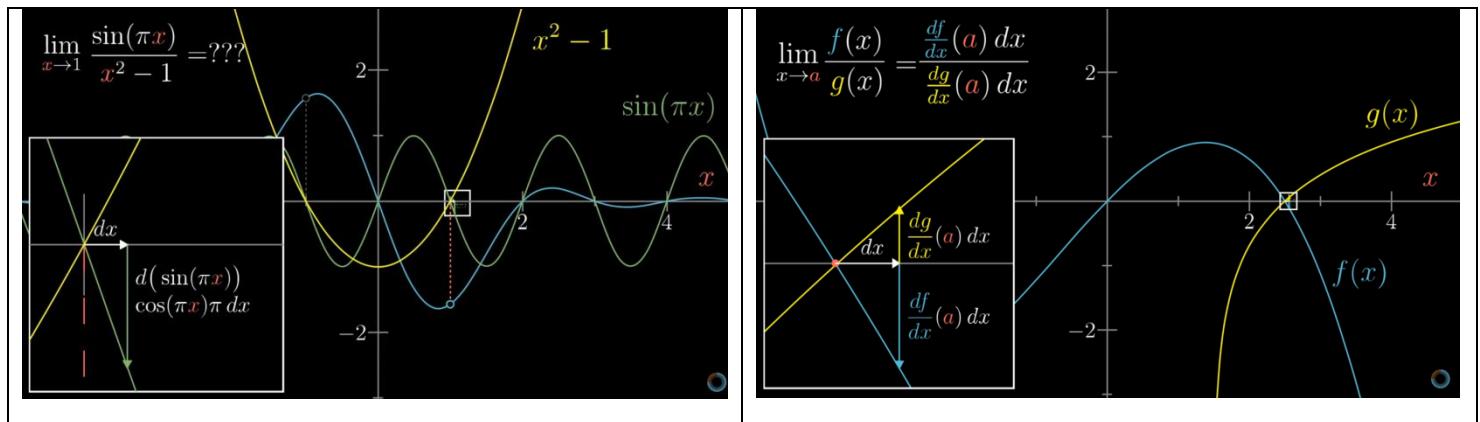
## Calculating limits

<div style="border: 1px solid #ccc; padding: 10px; background-color: #f0f0f0;"> <p>Indeterminate Forms:</p> <ul style="list-style-type: none"> <li><math>\frac{\pm\infty}{\pm\infty}</math></li> <li><math>\frac{0}{0}</math></li> <li><math>0 \cdot 0</math></li> <li><math>0^0</math></li> <li><math>1^{\pm\infty}</math></li> <li><math>\infty^0</math></li> <li><math>\infty - \infty</math></li> </ul> </div>	<p>A process to find limits as I see it is to begin with the traditional approach where the limit of a product of functions is the product of their limits. This is true for all functions except from the cases in which the product ends up to indeterminate forms. In these cases we must find other ways to calculate them and one of them is L'Hospital rule which works for cases in which the product results to <math>0/0</math>.</p>
--	---

## L'Hopital's rule

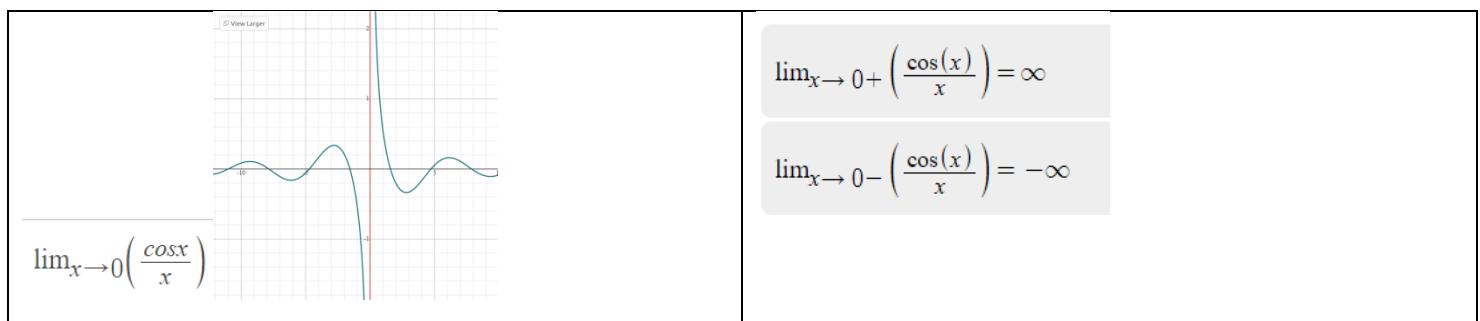
Only for 0/0

When a function is 0/0 (undefined) for a certain input, you can use a trick to calculate the value to which the function approaches as its input approaches that certain value. You define it as a ratio of two other functions and you want to calculate the limit of that ratio as their input approaches a certain value. To see what value do these functions have close to that input (lets say  $x=a$ ) we can see what happens in an input  $x+dx$ . If you can take the derivative of them at  $x=a$  (meaning that they are continuous at  $a$ , which means that if you zoom in close enough they look like straight lines) then their value is the value of their derivative at that point. So the limit can be replaced with a ratio of two derivatives evaluated at a certain input value.



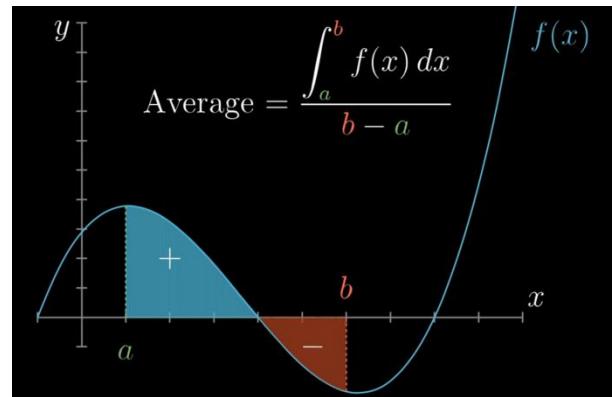
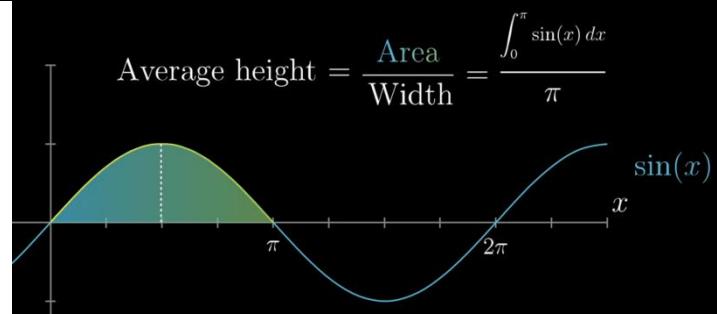
If  $\lim_{x \rightarrow a^-} f(x) \neq \lim_{x \rightarrow a^+} f(x)$  then the limit does not exist

$x \rightarrow a^-$  and  $x \rightarrow a^+$  is the way of saying  $x$  approaches  $a$  from the left and right respectively.



$$\lim_{x \rightarrow a} [f(x) \cdot g(x)] = \lim_{x \rightarrow a} f(x) \cdot \lim_{x \rightarrow a} g(x) \quad \text{With the exception of indeterminate form}$$

## Integrals

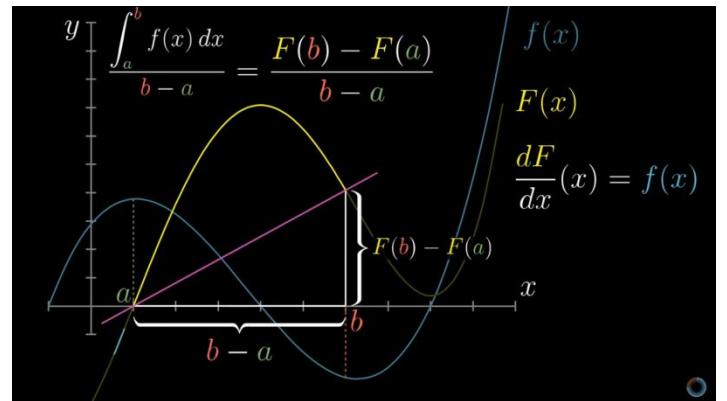


Find the average height between 0 and  $\pi$ .

The integral of the function between 0 and  $\pi$  divided by the width between 0 and  $\pi$ .

Or equivalently

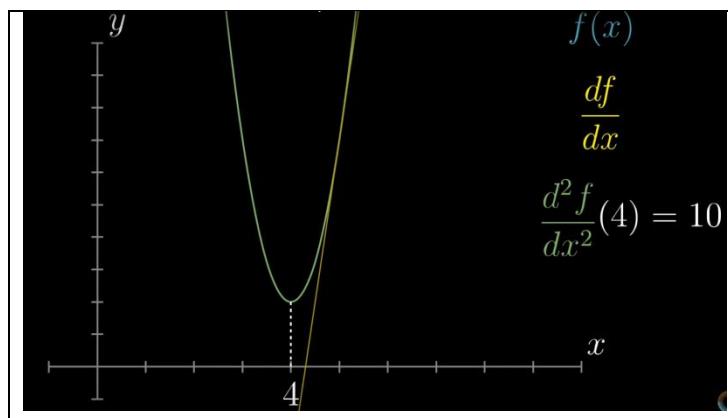
The average slope of tangents of the integral of the function between 0 and  $\pi$  which is the total slope between the start and end points.



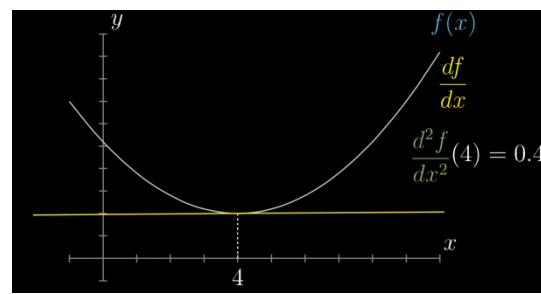
Capital F is the derivative of f.

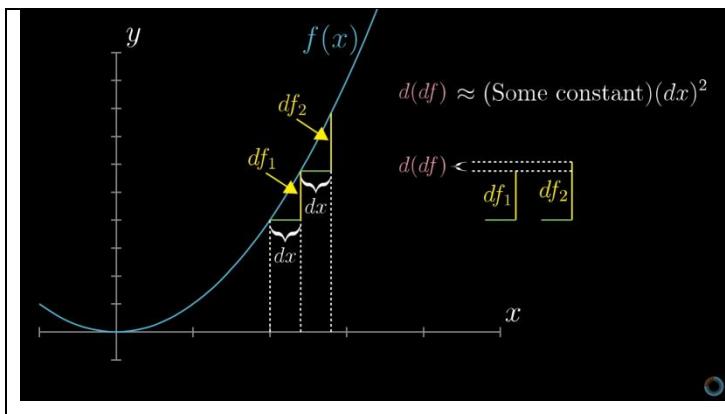
The notion of summing together infinite parts and dividing with infinity to get the average is represented by an integral divided by a length.

## Higher order derivatives



The slope of the tangent of  $f(x)$  increases rapidly as we move close to 4 so its derivative (the second derivative of  $f$ ) has a big value there in relation to a more flat curve





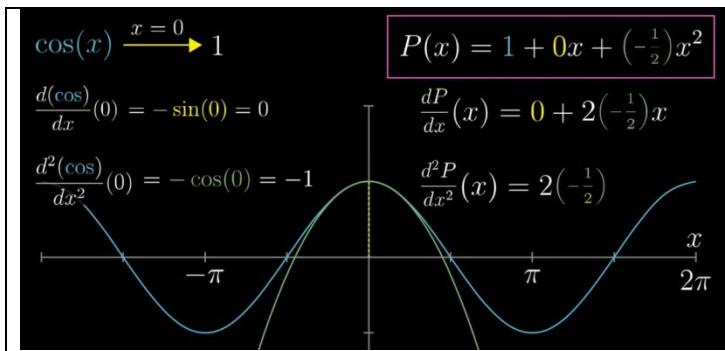
Graphical representation of the second derivative. The derivative is whatever the ratio of  $d(df)/dx^2$  approaches when  $dx$  approaches 0. the size of the change of the change divided by  $dx^2$

## Taylor series

One of the most useful things of higher order derivatives is that they help us approximate functions

Taylor polynomials are incredibly powerful for approximations, and Taylor series can give new ways to express functions. It is one of the most powerful tools mathematics has to offer for approximating functions.

The usefulness of Taylor series is that allow us to take non polynomial functions and approximate them with polynomials around some particular input. The reason this is important is that polynomials are much easier to deal with mathematically than other functions. They are easier to compute, to take derivatives from, to integrate etc.



For example lets think of the task to approximate the function  $\cos(x)$  for input values around 0, with a quadratic polynomial. The general formula for a quadratic polynomial is  $P(x)=c_0+c_1x+c_2x^2$ .

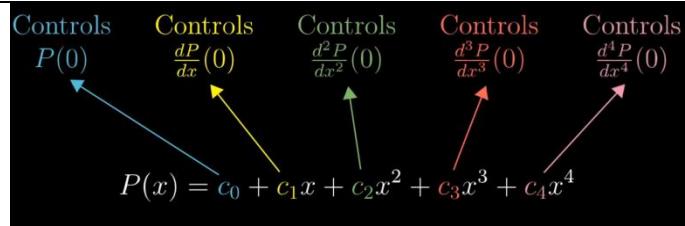
1. The polynomial must be 1 for  $x=0$  as  $\cos(x)$  does. From this equation we get  $c_0$ .
2. The slope of the tangent of the polynomial at  $x=0$  (its first derivative) must be equal with that of  $\cos(0)$ . from this we get  $c_1$ .
3. The slope of  $\cos(x)$  decreases around  $x=0$  (the second derivative of  $\cos(x)$  decreases) and specifically its value at 0 must be equal to the 2<sup>nd</sup> derivative of the polynomial (which means that the polynomial should curve similarly to the  $\cos(x)$  around 0). from this we get  $c_2$ .

In general, higher order derivatives of polynomial terms:

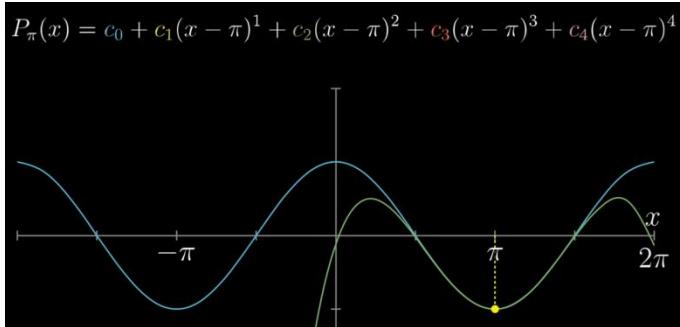
$$\frac{d^8}{dx^8}(c_8 x^8) = \underbrace{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8}_{8!} c_8$$

This way we have calculated a quadratic polynomial that approximates 0. We can make better approximations if we add more terms to the polynomial. We find its term using the respective high order derivative. In this case for cubic term  $c_3=0$  which means that the specific quadratic polynomial is not only the best quadratic approximation but it is also the best cubic approximation.

The higher order derivatives of the function that we want to approximate, match the derivatives of the polynomial. So the polynomial is constructed only with information of higher order derivatives at a specific point!

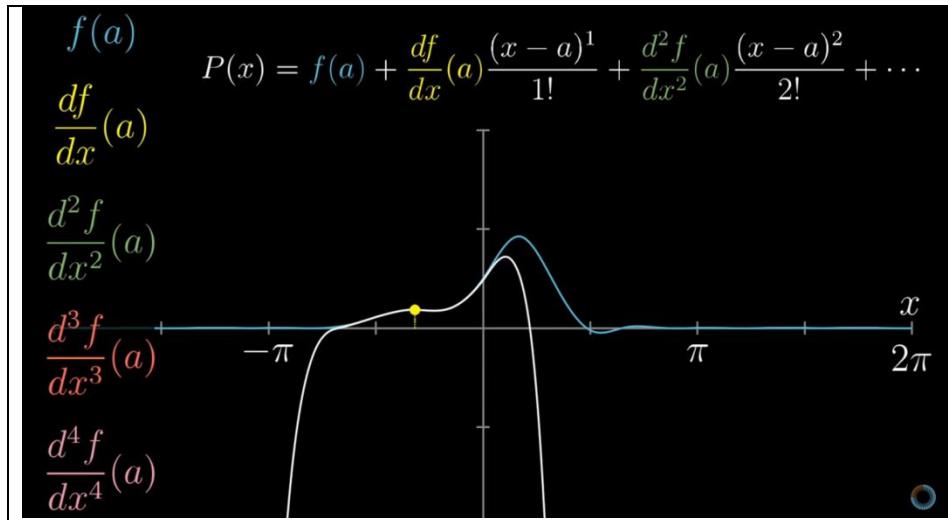


Adding new higher power terms doesn't affect the lower terms the coefficient of which are already calculated. This happens since the second derivative of the polynomial evaluated in 0, all initial terms with higher order than 2, leave an x term or higher in the derivative, but they become 0 for  $x=0$ . This means that the constant of the  $(x)^2$  term is the only constant that affects the second order derivative of the polynomial.



This is true for finding an approximation around 0.

If you want to find an approximation around  $\pi$ , then you should write the polynomial in terms of  $x-\pi$  so that when you calculate the second derivative in  $\pi$  all terms with higher order than 2 leave an  $x-\pi$  term or higher in the derivative, but they become 0 for  $x=\pi$ . This means that the constant of the  $(x-\pi)^2$  is the only constant that affects the second order derivative of the polynomial.



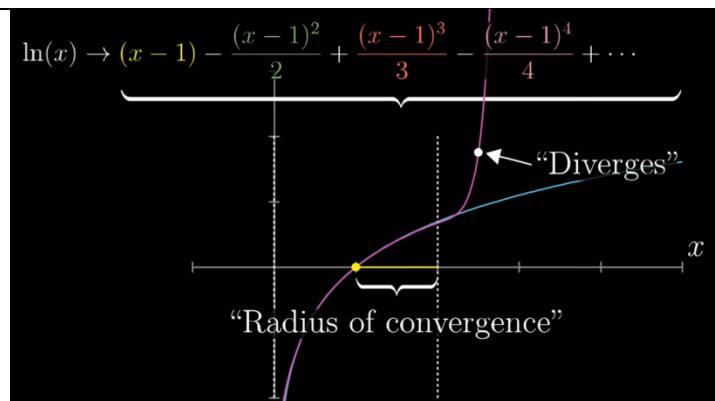
The  $P(x)$  is called a Taylor polynomial.

This has tremendous applications in many physics and engineering problems for which an angle theta slightly oscillates around a certain value.

"Converges" to  $\frac{1}{2}$

$$\frac{1}{3} + \frac{1}{9} + \frac{1}{27} + \frac{1}{81} + \frac{1}{243} + \dots = \frac{1}{2}$$

In math the sum of infinitely many terms is called a series, so if you take into account all the Taylor polynomial terms it becomes a Taylor series. If you have a series where you add more and more terms gets you increasingly close to a certain value then we say that the series converges to that value and we can write an "equality" for it where the series equals the value that it converges to.



A Taylor series could converge to a certain value (the value of the function that we want to approximate at a certain input) only for a certain range of inputs. This range is called the radius of convergence for the Taylor series. The series diverges outside of that range.

We can say that the effect of the derivatives of the function at that input doesn't propagate outside of a certain range.

There are test to see if a certain Taylor series converges or not.

I get this "If you don't plan to do anything bad you have nothing to fear" narrative. It really has a point in a prosperous world. What troubles me though is that no one can guarantee that our world will continue to prosper without any setbacks in between. In rough times populists prosper and its they that will be deciding what is good and what is bad. This time though they would have the means to

## Misc

### differentiation methods

1. symbolic
2. numerical
3. automatic

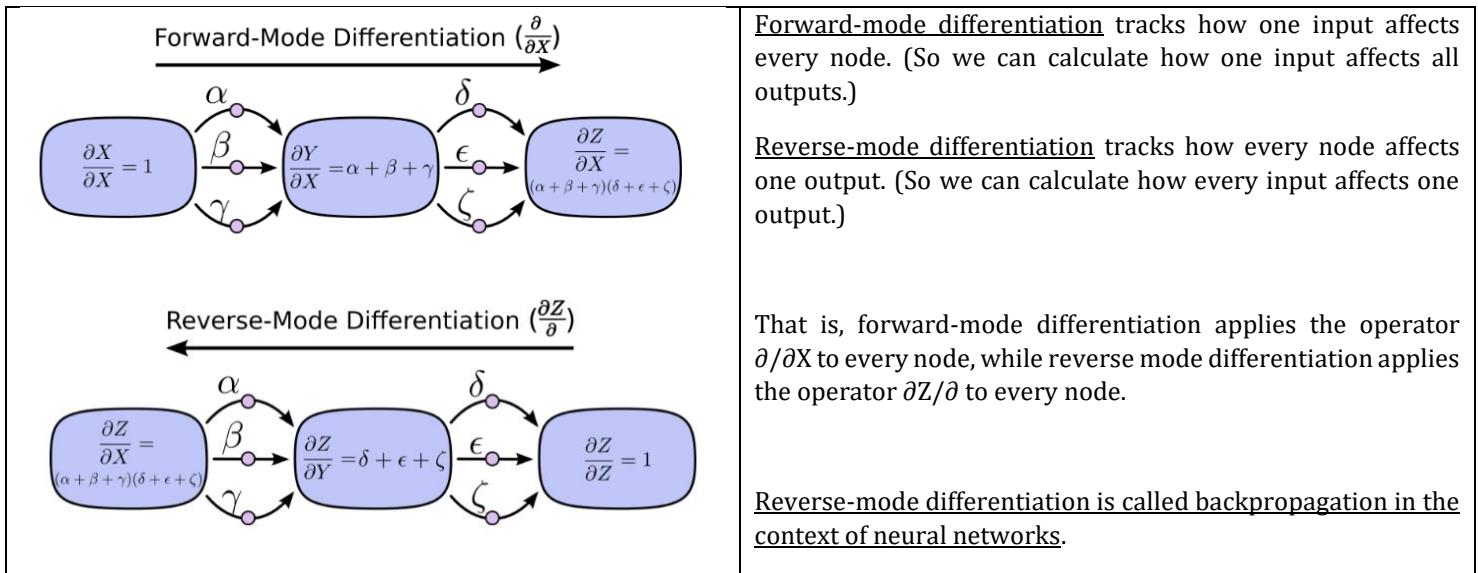
- **Automatic differentiation**

<http://colah.github.io/posts/2015-08-Backprop/>

It is a set of techniques to evaluate the derivative of a function specified by a computer program. AD exploits the fact that every computer program, no matter how complicated, executes a sequence of elementary arithmetic operations (addition, subtraction, multiplication, division, etc.) and elementary functions ( $\exp$ ,  $\log$ ,  $\sin$ ,  $\cos$ , etc.). By applying the chain rule repeatedly to these operations, derivatives of arbitrary order can be computed automatically, accurately to working precision, and using at most a small constant factor more arithmetic operations than the original program.

On simple terms, it is a way of automatically computing the derivatives of the output of a function using chain rule. Almost every function can be computed as a composition of simple functions which have simple derivatives. Using this fact, you can compute the derivative of any function that can be written as composition of simpler functions.

Both of the classical methods (numerical and analytical) have problems with calculating higher derivatives, where complexity and errors increase. Finally, both of these classical methods are slow at computing partial derivatives of a function with respect to many inputs, as is needed for gradient-based optimization algorithms. Automatic differentiation solves all of these problems.



(Are there any cases where forward-mode differentiation makes more sense? Yes, there are! Where the reverse-mode gives the derivatives of one output with respect to all inputs, the forward-mode gives us the derivatives of all outputs with respect to one input. If one has a function with lots of outputs, forward-mode differentiation can be much, much, much faster.)

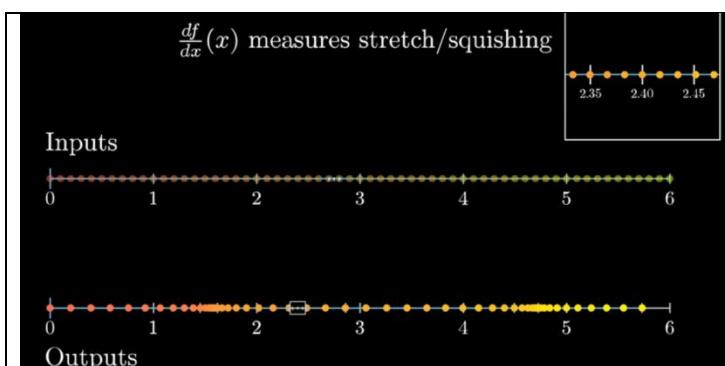
- **Numerical differentiation**

Numerical differentiation (the method of finite differences) can introduce round-off errors in the discretization process and cancellation.

- **Symbolic differentiation**

Analytical or symbolic differentiation faces the difficulty of converting a computer program into a single mathematical expression and can lead to inefficient code. ... Although computer algebra could be considered a subfield of scientific computing, they are generally considered as distinct fields because scientific computing is usually based on numerical computation with approximate floating point numbers, while symbolic computation emphasizes exact computation with expressions containing variables that have no given value and are manipulated as symbols.

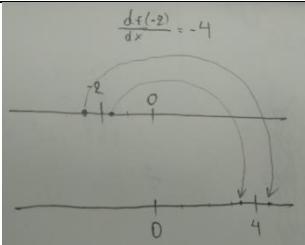
## Transformational understanding of derivatives



Another useful way to visualize derivatives is in the context of transformations. You have two number lines and a function  $f(x)$ . In the first number line you mark the inputs to the function and in the second the outputs. The size of the stretching or squishing is the derivative of the function  $f$ .

When they are stretched out a lot, then a small change in input causes a big change in the output. In the context of graphs, this is represented by a big slope of the graph at that point.

For example for  $f(x)=x^2$  for  $x=2$  the points around 2 are stretched by a factor of 4, and for  $x=3$  they are stretched by a factor of 6. This because  $df/dx=2x$



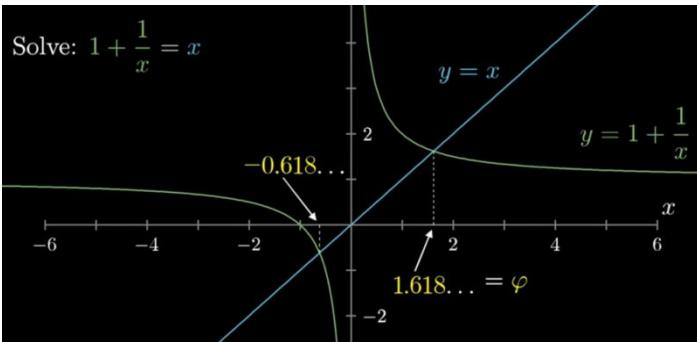
If the derivative is smaller than 1 then it means that the stretching factor is smaller than 1 which means that the values are actually squished. For a derivative at a certain input and value of 1/2 they are squished in half the length.

If the derivative is negative (for example  $df(-2)/dx=-4$ ) the output values don't only stretch out, they also flip around for example -2.1 goes to 4.2 while -1.9 goes to 3.8

## Infinite expressions and fixed points

$$1 + \frac{1}{1 + \dots}}}}}}} = x$$

$$1 + \frac{1}{x} = x$$



$$1 + \frac{1}{1 + \dots}}}}}}} = c = -0.65 = -0.650\dots$$

$$f(c) = 1 + \frac{1}{-0.65} = -0.538\dots$$

$$f(f(c)) = 1 + \frac{1}{1 + \frac{1}{-0.538\dots}} = -0.857\dots$$

$$f(f(f(c))) = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{-0.857\dots}}} = -0.167\dots$$

$$f(f(f(f(c)))) = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{-0.167\dots}}}} = -5.000\dots$$

???

Imagine that we want to calculate what value this infinite ratio approaches. A way to think about it, is to say that this whole ratio equals to an unknown value  $x$ , to which it converges to. But you can clearly see that the ratio contains  $x$  inside it and consequently it can be represented as  $1+1/x=x$ . Plotting the two functions  $f(x)=1+1/x$  and  $f(x)=x$  we can see that there are two solutions,  $x=\varphi$  and  $x=-1/\varphi$ . But the same ratio can't converge to two different values, only to one, but which one and why?

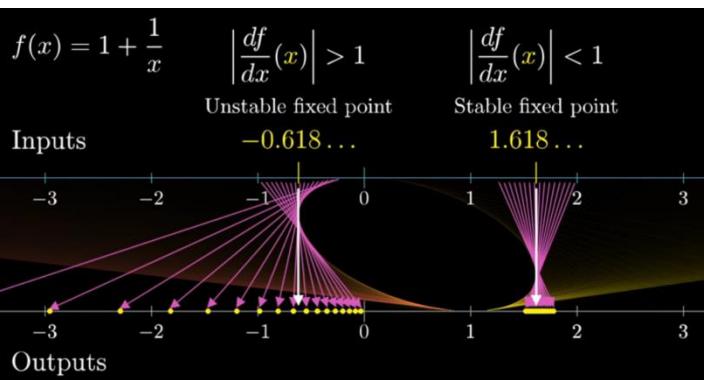
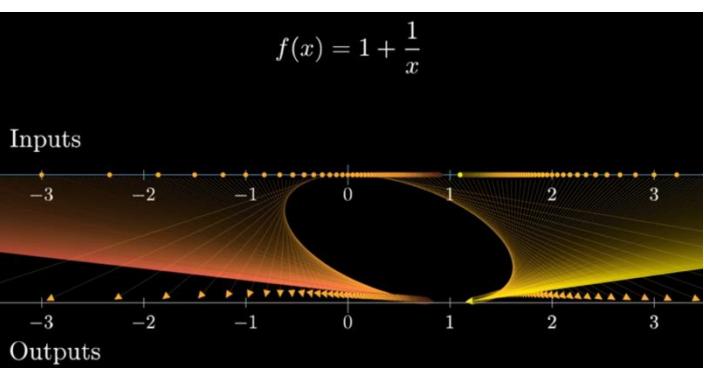
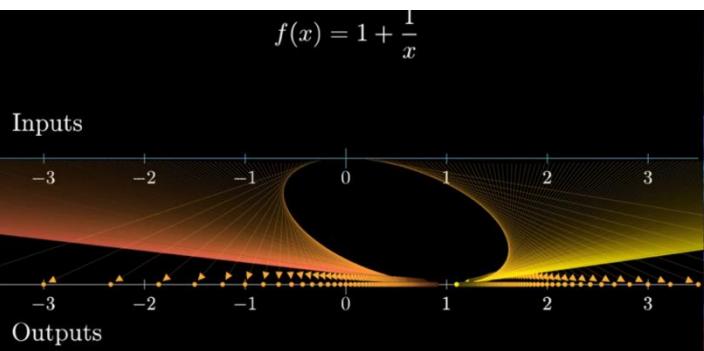
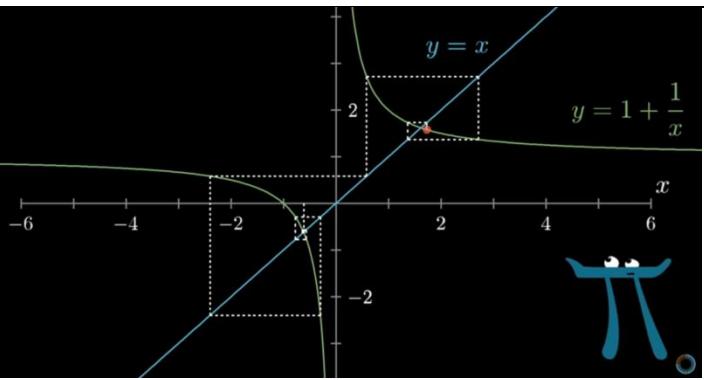
A way to approach the solution to the *infinite fraction* is to think of having a function  $f(x)=1+1/x$  and applying it over and over again to a random input and see where the result converges to. Then repeat for another random input value and compare, then another and so one. Doing it, we will see that all inputs converge to  $\varphi$ . Some of them might wander around  $-1/\varphi$  for a little bit, but then they move towards  $\varphi$ . We say that  $\varphi$  is the fixed point of the

Notice that if you start from  $-1/\varphi$  then you stay fixed in that value. No matter how many times you apply  $f(x)$  you remain in that point. This is the only point for which the ratio "converges" to  $-1/\varphi$ . Even if you start very close to it, you end up in  $\varphi$ .

The graph understanding

The reason this happens, has to do with the derivative of the function  $1+1/x$ . Both the graph understanding of derivatives and the transformational understanding of derivatives can help us understand this behaviour.

We start from a random  $x$  value, we go to its  $f(x)$  then we move horizontally to cross the  $y=x$  line where the  $y$  value (the output of  $f(x)$ ) is the same with the  $x$  value, so we can use this  $x$  value as the new input to  $f(x)$  and repeat. Doing it we will see that all



values expect  $-1/\varphi$  end up to  $\varphi$ . And this has to do with the slope of  $f(x)$ .

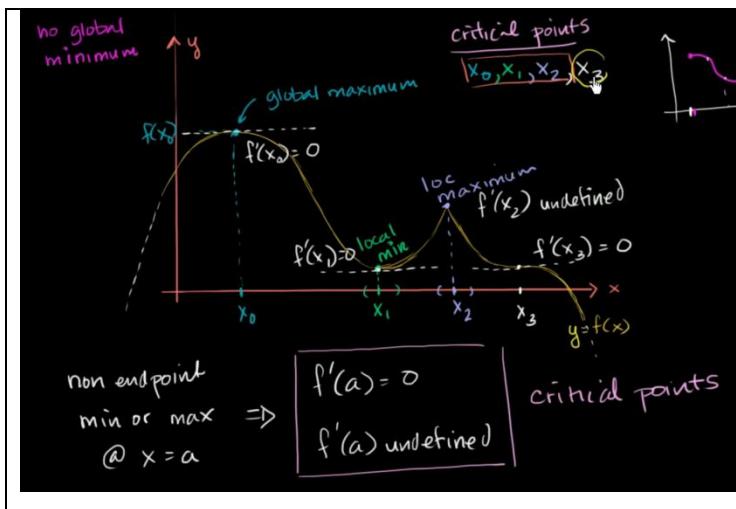
In the transformation representation of derivatives repeatedly applying the same function means taking the values of the output number line, moving them to the input line and getting the new output line. Then repeat. One advantage of this approach is that you can see the behaviour of many inputs at once (like beginning from many inputs in the graph representation). The derivative of  $f(\varphi)$  is negative and smaller than one, so input values around  $\varphi$  gravitate around  $\varphi$  in the output, while the derivative in  $-1/\varphi$  is  $-2.8$  which means that input values around  $-1/\varphi$  are repelled away from there in the output. This explains why values are moving away of  $-1/\varphi$  and end up in the region of  $\varphi$ .

In  $\varphi$  you can think of the input dots turning around the output  $\varphi$  and closing in on it as if it was a gravitational point while  $-1/\varphi$  acts like an anti-gravitational point throwing the points away. The first one is a **stable fixed point** while the second is an **unstable fixed point** of the function  $f(x)=1+1/x$ . As I understand it, fixed points are the values to which the process of repeatedly applying a function converges to.

So the stability of a fixed point is determined by whether the magnitude of the derivative of the function evaluated to the point is greater or smaller than 1.

So if you think of this infinite fraction through a limiting process (in which you want to approach the value that it converges to) then the most valid answer is  $\varphi$ .

## Finding Extrema



- Critical points of a function are all the points for which the first derivative is 0 or undefined.
- Not all critical points are extrema of the function
- If the first derivative has a different sign before and after the critical point then this point is an extrema
- If the first derivative is  $<0$  before the critical point and  $>0$  after it, then the point is minimum. In other words if the second derivative at the point is positive (the first derivative increases around the point)

## Misc

Calculus is the mathematics of change, of everything that changes.

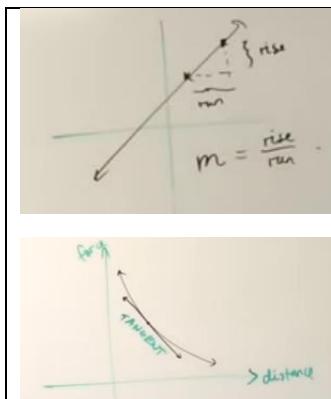
The real name is infinitesimal calculus

Invented simultaneously by Newton and Leibniz

Algebra is the math of relationships

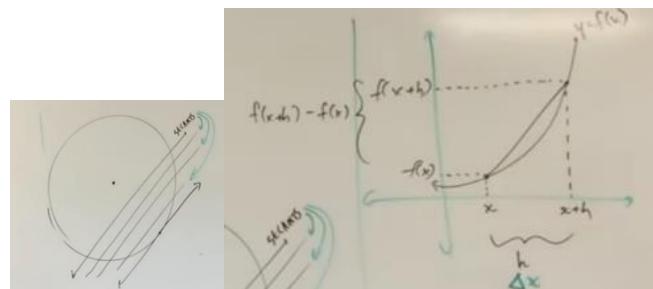
Probability or Statistics is the math of uncertainty, chance

Geometry is the math of space, shapes.



In general gradient measures how something changes. For example, in a line you can take two points and you measure how much it changes in one quantity, versus how much it changes in the other. So gradient ( $m$ ) is the ratio of **rise to run**.

But you can't do this same thing in a function that is not linear since any two points give different result. The gradient constantly changes. What Newton did when studying the nature of gravity, was to try to calculate the gradient of the tangent. The problem is that the tangent is by definition tangent at one point only, so you can't use rise/run since you need two points for that. What they did, was to approach the value of the gradient of the tangent by calculating the gradient of secants as the length of the secant approaches to 0.



$$m_{\text{secant}} = \frac{\text{rise}}{\text{run}} = \frac{f(x+h) - f(x)}{h}$$

**FIRST PRINCIPLES**

$$m_{\text{TANGENT}} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \frac{dy}{dx}$$

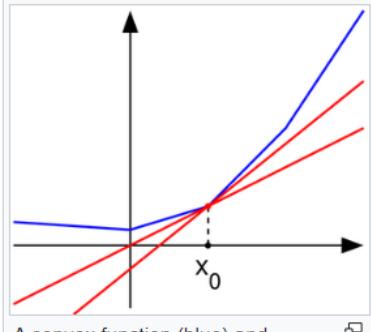
(Gradient Function)

$Dy/dx$  is the ratio of change of  $y$  to change of  $x$  as  $h$  (or  $\Delta x$ ) goes to 0.

Limits. You make conclusions about something that you don't know, based on the things you know.

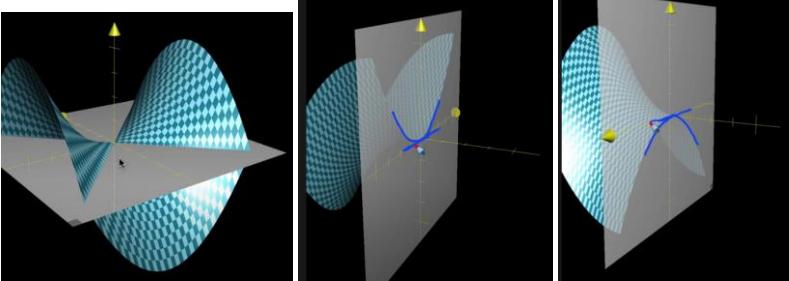
## Subgradients

In mathematics, the subderivative, subgradient, and subdifferential generalize the derivative to convex functions which are not necessarily differentiable.

 <p>A convex function (blue) and "subtangent lines" at <math>x_0</math> (red).</p>	<p>for any <math>x_0</math> in the domain of the function one can draw a line which goes through the point <math>(x_0, f(x_0))</math> and which is everywhere either touching or below the graph of <math>f</math>. The slope of such a line is called a subderivative (because the line is under the graph of <math>f</math>).</p> <p>The definition of the derivative is extended to include the set of all possible gradients at a non-differentiable point. The point at <math>(0,0)</math> is not differentiable, but it is subdifferentiable. We can take the set of all lines that touch at just that point, and that is the subderivative/gradient.</p>
---	---

## Saddle points

[https://www.youtube.com/watch?v=8aAU4r\\_pUUU](https://www.youtube.com/watch?v=8aAU4r_pUUU)

	<p>multivariate functions can have saddle points points at which the tangent (derivative) is 0 but they are neither minima nor maxima.  For example <math>f(x,y)=x^2-y^2</math></p>
--	---