AI Summer Camp

Elsiddig Elsiddig

King Abdullah University of Science and Technology, Saudi Arabia

July 5, 2023

Calculus Agenda

- Functions
- Differentiation
- Maxima and Minima
- Line of Best Fit
- Multivariate Functions

Relation to AI?

In a reductionist view AI boils down to optimizing a loss function.

Loss Function: The loss function is a measure for how good the model performs. A Loss function is either designed from collected data or is made specific to the problem at hand.

Optimizing Algorithm: Many optimizing algorithms utilize derivatives of the loss function. Optimizing means minimizing in most contexts

Learning: The terminology learning can be thought of as the process of applying an iterative optimizing algorithm to a minimize a given loss function.

Suppose we have data of the form (X, Y). We may be interested in approximating a function f that satisfies $Y = f(X)$.

Neural Networks: Consecutive composition of linear maps and a non-linear map.

Functions

Definition

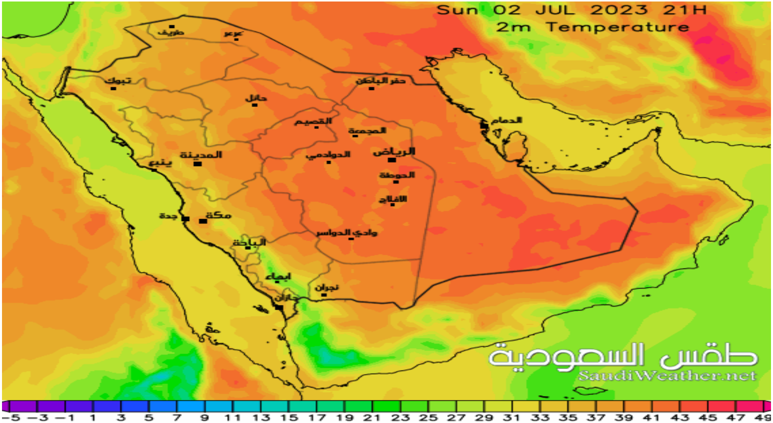
A function (mapping) is a relation, law or rule from a set of inputs into a set of possible outputs where each input has a unique output.

Example

The polynomial $p : \mathbb{R} \to \mathbb{R}$ defined by $p(x) = 5x^3 - 2x^2 + 17$.

Exercise: evaluate $p(0)$ and $p(-1)$.

Temperature measurements can be thought of as a function that takes a point in land as input (ignoring factors such as wind and humidity).



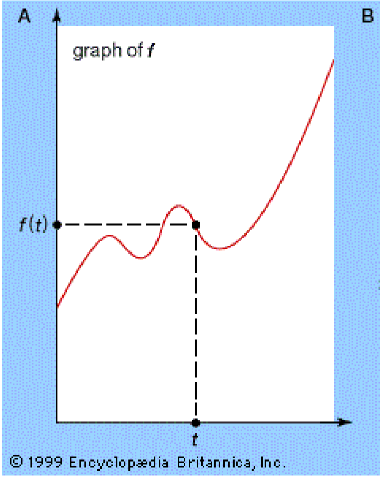
Functions are objects which return the same value whenever given the same input.

Functions can also take "several" inputs.

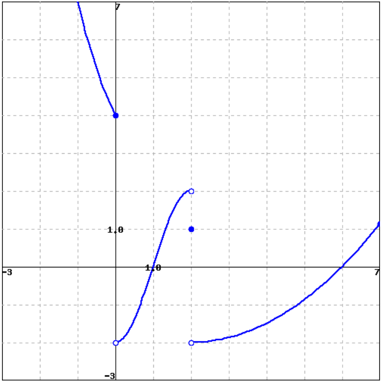
Example

Area of rectangle. Height and width are input and area$=$ height$\times$width is the output.

Graph of a function



(a) label 1



(b) label 2

Figure: Examples of functions

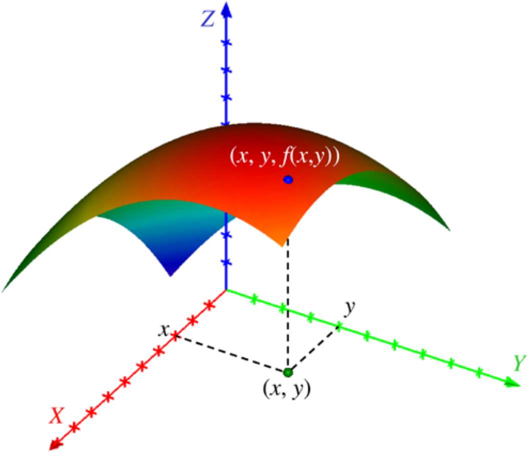
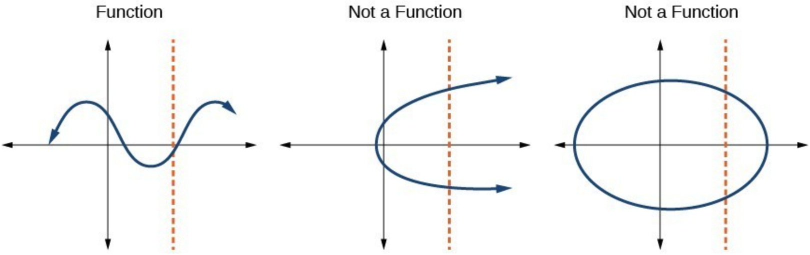


Figure: function graph in 3d

Vertical Line Test

A curve is a graph of function if every vertical line cuts the curve once at most.



Domain and Co-Domain

Definition

The domain of a function is the set of allowed inputs for that function.
The co-domain is the set of possible outputs.

The function $f(x) = \sqrt{x-1}$ has domain $[1, \infty)$.
The polynomial p defined previously works in all real numbers $\mathbb{R}$.

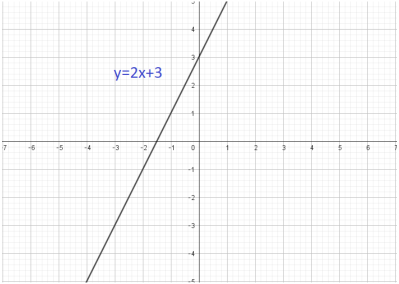
The area of rectangle function has pairs of positive real numbers as inputs (written $\mathbb{R}^+ \times \mathbb{R}^+$).

Exercise: What is the domain of the function $f(x) = \log(x^3 + 1)$?

Important Examples of Functions

Linear Maps

$$f(x) = mx + b$$

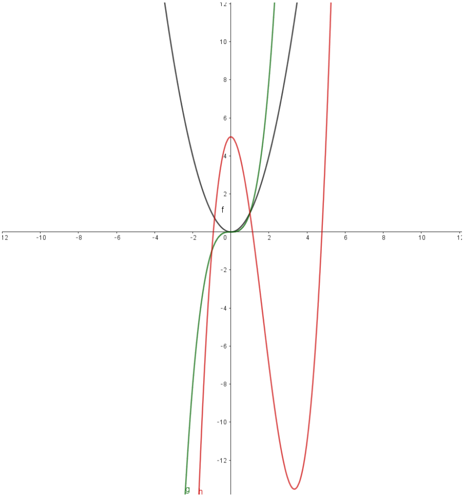


m is called the slope and b is called the intercept.

Exercise: plot the line $y = \frac{1}{2}x + 3$.

Polynomials

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$$



Exponential Map

$$f(x) = e^x$$
$$f(x) = 2^x$$

Trigonometric Functions

$$f(x) = \sin(x)$$
$$f(x) = \cos(x)$$

Logarithmic Functions

$$f(x) = \log_b(x)$$

Composition

Definition

Let $f : A \to B$ and $g : B \to C$ be two functions. We define the composition $g \circ f : A \to C$ as

$$(g \circ f)(x) = g(f(x))$$

Example: $f(x) = x^2$ and $g(x) = x - 1$. Then

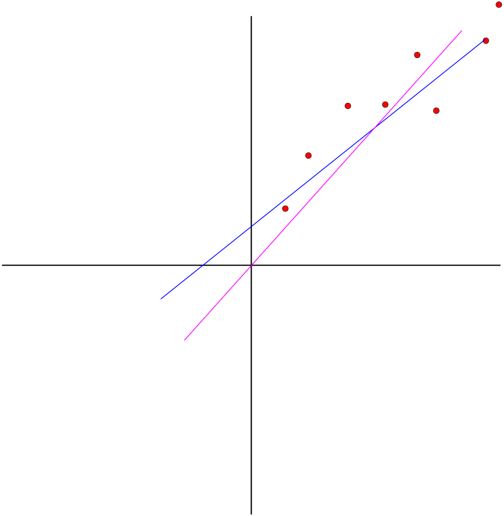
$$(g \circ f)(x) = (x - 1)^2$$

while

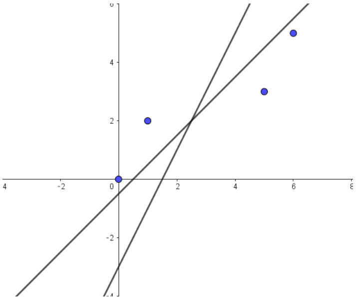
$$(g \circ f)(x) = x^2 - 1$$

Loss Function: Linear Regression Example

Which line is a better fit for the red data point?



Let's look at a simpler example. Suppose we have the following data points $(0,0),(1,2),(5,3),(6,5)$. We would like to fit a line to these data points.



Two lines $y = x - 0.5$ and $y = 2x - 3$ are plotted (which equation corresponds to each line?). We want to ask which one of them is a better fit for the data?

To asses the fit we calculate the sums squared errors. This quantity is calculated by subtracting the prediction of the model from the actual data and summing their squares.

In our case our inputs are $0, 1, 5, 6$ and the outputs are $0, 2, 3, 5$. To assest the fit of the line $y = x - 0.5$ we calculate its outputs given the same previous inputs. The outputs of the line are $-0.5, 0.5, 4.5, 5.5$. We now calculate

$$(-0.5 - 0)^2 + (0.5 - 1)^2 + (4.5 - 3)^2 + (5.5 - 6)^2 = 3$$

We do the same for the line $y = 2x - 3$. The outputs are $-3, -1, 7, 9$. We have

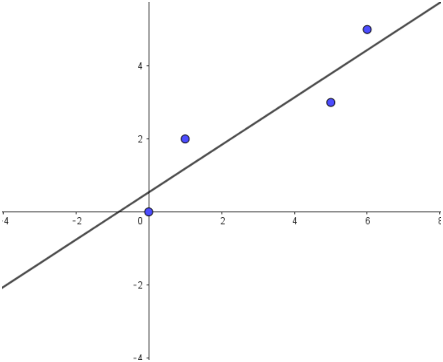
$$(-3 - 0)^2 + (-1 - 1)^2 + (7 - 3)^2 + (9 - 6)^2 = 38$$

Thus the line $y = x - 0.5$ is a better fit for the data. However, this line is not the best.

The best line is

$$y = 0.65x + 0.54$$

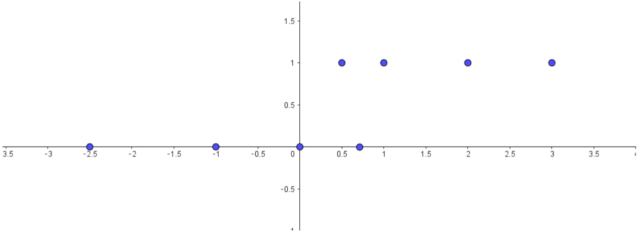
whose loss function value is 1.9.



How do we find this line? Finding the line that minimizes the sum of squared errors is an example of least squares method.

Logistic Regression Example

Suppose we have the follow data points. Such problems where the response (output) is either 0 or 1 are called binary classification problems.

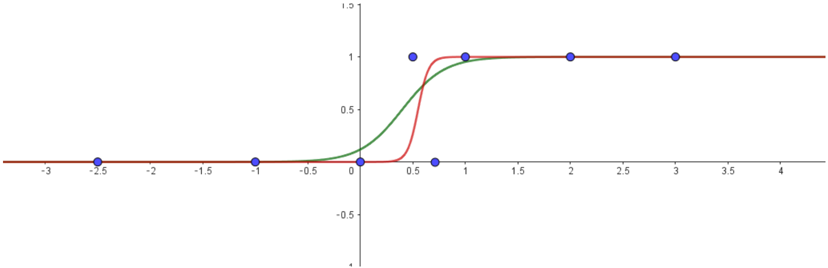


We would like to calculate a threshold at which the model will transit from 0 to 1.

Which model? For this problem instead of fitting a line we will fit (find m and b) the function $\sigma(mx+b)$ where

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

In the exercises σ was shown to be increasing and bounded between 0 and 1.



Plotted $\sigma(5x-2)$ and $\sigma(20x-11)$ (which is which?). What is our loss function and how do we find the best values for m and b for the chosen loss function?

Goal: Minimizing Functions

Differentiation

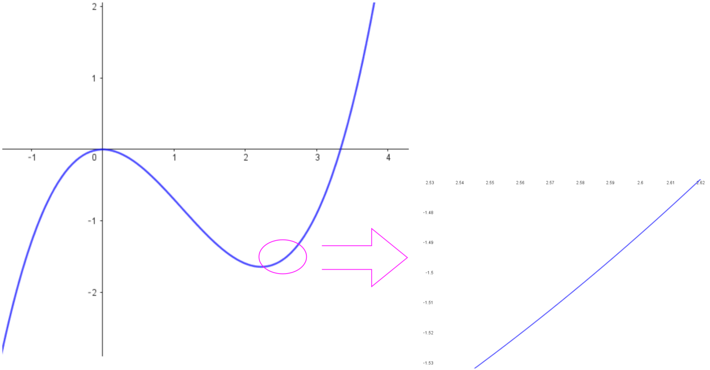


Figure: Linear approximation

Many of the function you have encountered have the property that when we select a point and keep zooming around that point the curve will start look like straight. Such functions are called differentiable functions.

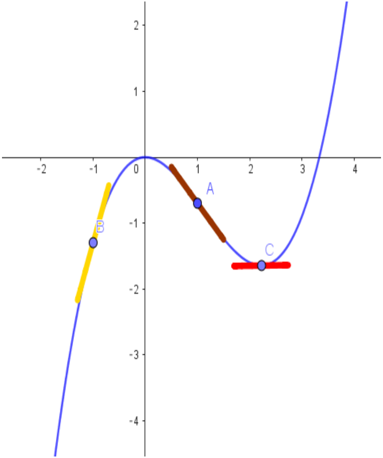
Example

All the examples shown previously are differentiable on their domains.

- The zooming process described previously means that the function has a well-defined tangent at the point being zoomed around.
- The zooming process does not end, we can always zoom in. However, after a while the linear trend becomes clear and the appearance of the zoomed curve does change except marginally. The result of this process is called a limit.
- The tangent at a point x is a line. We only need its slope to write its equation (because we already the tangent passes through x, clear?).
- The tangent at x is the "best" line that approximates the function in a small vicinity of x. This is why it is also called the linear approximation at x.

Why Tangents?

Tangents show us the local behaviour of the function.



In the figure below the tangent at point A has a negative slope and the function is decreasing. At point B the tangent slope is positive and the function is increasing. Point C is called critical point.

All local maxima and minima are critical points.

Definition

The derivative of a function f at a point x, denoted by $f'(x)$, is the slope of the tangent at point x. Formally defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

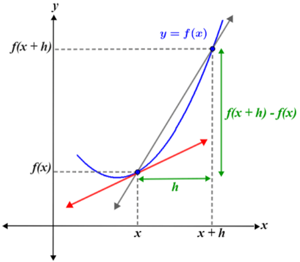


Figure: The value of h is exaggerated. The red line is the tangent and the black line is an approximation of the tangent. When h becomes tiny the black line will converge to the red line

Calculation of Derivative

Constant functions: $f(x) = c$.
The graphs of constant functions are horizontal lines. Since the function is already a line, no need to approximate it by a line. Hence the derivative is the slope of the line itself which is 0. We write

$$f'(x) = 0$$

Linear functions: $f(x) = mx + b$.
The same reasoning above applies to linear functions. We have

$$f'(x) = m$$

Formally for constant functions $f(x) = c$ we write

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{c - c}{h} = \lim_{h \to 0} 0 = 0$$

For linear functions $f(x) = mx + b$ we write

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{m(x+h) + b - (mx+b)}{h} = m$$

Notice that because the functions are already lines the lim does not play a real significant role.

Quadratic: $f(x) = x^2$

$$
\begin{aligned}
f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} \\
&= \lim_{h \to 0} \frac{2xh + h^2}{h} \\
&= \lim_{h \to 0} 2x + h \\
&= 2x
\end{aligned}
$$

Derivatives of fundamental functions

- $(c)' = 0$
- $(mx + b)' = m$
- $(x^n)' = nx^{n-1}$
- $(\sin(x))' = \cos(x)$
- $(\cos(x))' = -\sin(x)$
- $(e^x)' = e^x$
- $(\ln(x))' = \frac{1}{x}$
- $(2^x)' = 2^x \ln(2)$

Differentiation Properties

Linearity

$$(f+g)' = f' + g'$$

$$(cf)' = cf' \qquad , c \in \mathbb{R}$$

Let $f(x) = x^2 + \sin(x) - 5\ln(x)$. The derivative of f is

$$f'(x) = 2x + \cos(x) - \frac{5}{x}$$

Product Rule

$$(fg)' = f'g + fg'$$

Division Rule

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

Example:

$$(x\sin(x))' = \sin(x) + x\cos(x)$$

$$(\tan(x))' = \left(\frac{\sin(x)}{\cos(x)}\right)' = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)} = \frac{1}{\cos^2(x)} = \sec^2(x)$$

Chain Rule

$$(g \circ f)'(x) = f'(x)g'(f(x))$$

Examples:

$$(\cos(\sin(x)))' = -\cos(x)\sin(\sin(x))$$

$$\frac{d}{dx}(x^2 + 5)^5 = 10x(x^2 + 5)^4$$

Note: $\frac{d}{dx}$ is just a alternative notation for the derivative.

Behaviour of a function from its derivative

Figure: Red: function. Blue: derivative

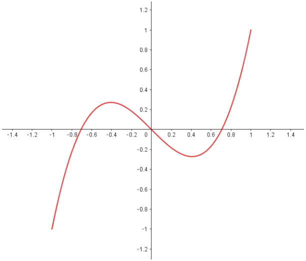
$$f'(x) > 0 \Rightarrow f \text{ is increasing near } x$$

$$f'(x) < 0 \Rightarrow f \text{ is decreasing near } x$$

$$f'(x) = 0 \Rightarrow x \text{ is called a critical point}$$

Finding Extrema

Suppose we want to find the maximum and minimum of the function $f(x) = 2x^3 - x$ in the interval $[-1, 1]$.
From the plot below we see that the maximum is at $x = 1$ and the minimum is at $x = -1$ which are the boundaries of th interval.



However, if we consider the same function of the the interval $[-0.7, 0.7]$ the maximum and minimum are not boundaries of the interval.

Standard Proceedure to find extrema

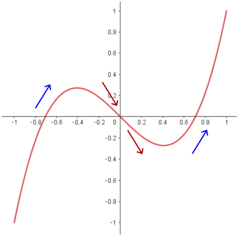
- Find the critical points and evaluate the function at those *point*.
- evaluate the function at the boundaries of your interval.
- The smallest of these evaluations is the minimum and the largest is the maximum.

We apply the procedure to the function $f(x) = 2x^3 - x$ on the interval $[-1, 1]$.
The derivative is $f'(x) = 6x^2 - 1$. To find the critical points we solve $6x^2 - 1 = 0$. The solutions are $x = \pm\frac{1}{\sqrt{6}}$.
The candidates for the minimum and maximum are
$f\left(\frac{1}{\sqrt{6}}\right) = -\frac{2}{3\sqrt{6}}, f\left(-\frac{1}{\sqrt{6}}\right) = \frac{2}{3\sqrt{6}}, f(1) = 1, f(-1) = -1$

Therefore the maximum is -1 and the maximum is 1.

Testing critical point type

Suppose x is a critical point. If the derivative is negative just before x and positive after it. The point x must be a local minimum. This is called the first derivative test.



We expect the derivative to be increasing near x. That is $f''(x) > 0$. This is called the second derivative test.

What are the corresponding tests for the maximum? Keep in mind that some points can be critical but neither maximum nor minimum.

Solving our Linear Regression Problem

In slide 17 we had the points $(0,0),(1,2),(5,3),(6,5)$ and we wanted to fit a line to them. We will now show how to find the best line of fit shown in slide 18.

A line equation is of the form $mx+b$ and we seek the best m and b. The evaluations at $x=0,1,5,6$ are $b, m+b, 5m+b, 6m+b$. Hence the loss function is

$$(b-0)^2+(m+b-2)^2+(5m+b-3)^2+(6m+b-5)^2$$

We will denote it $\mathrm{loss}(m,b)$. We would like to find the minimum of $\mathrm{loss}(m,b)$.

Although we have two variables in $\mathrm{loss}(m, b)$ the same logic for finding the minimum applies. We differentiate w.r.t m and b and equate to 0.

$$\frac{\partial}{\partial m}\mathrm{loss}(m, b) = 124m + 24b - 94$$

$$\frac{\partial}{\partial b}\mathrm{loss}(m, b) = 24m + 8b - 20$$

Equating with 0 we have the system

$$\begin{cases} 124m + 24b - 94 = 0 \\ 24m + 8b - 20 = 0 \end{cases}$$

The solution of this system is $m = \frac{17}{65} \approx 0.65$ and $b = \frac{7}{13} \approx 0.54$.

We have to check the boundaries. When wither m or b go to ∞ the function $\mathrm{loss}(m, b)$ grows to ∞ as well.

Therefore we conclude that the critical point $(0.65, 0.54)$ is a minimum and the line that best fits the data is $0.65x + 0.54$.