IN3050 Mathgroup, Derivatives

Tobias Opsahl

March 9th, 2023

UNIVERSITY OF OSLO

Outline

1. Notation
2. Computations
3. How we do it (at blackboard)

What is the derivative

- **Most importantly:** A derivative (or gradient) of a function is a measure over how *fast that function changes*.

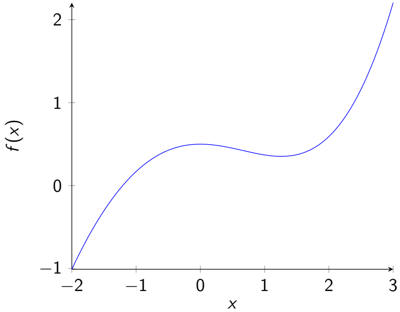
What is the derivative

- **Most importantly:** A derivative (or gradient) of a function is a measure over how *fast that function changes*.
- This can be precisely formulated mathematically, and calculated many ways (analytically, numerically, automatic differentiation, etc).

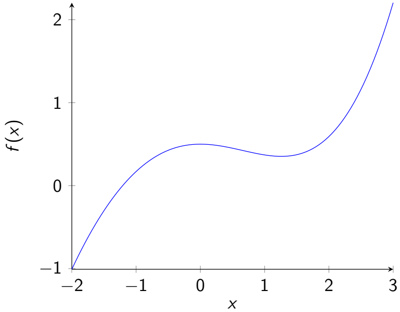
What is the derivative

- **Most importantly:** A derivative (or gradient) of a function is a measure over how *fast that function changes*.
- This can be precisely formulated mathematically, and calculated many ways (analytically, numerically, automatic differentiation, etc).
- In machine learning, we are often interested in the gradient of the *loss*-functions with respect to its *weights*. That means, how much does the loss-function change when we change the weights a little bit.

The Derivative Graphically

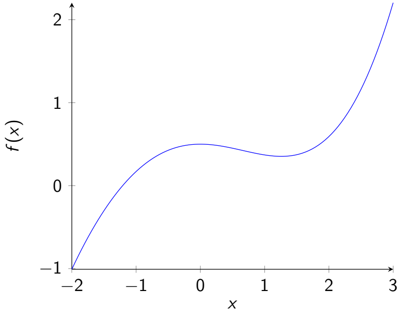


The Derivative Graphically



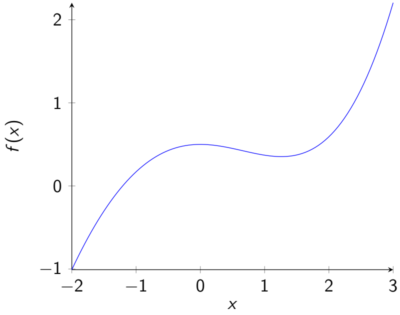
- At the region between -2 and 0, the function increases, so the derivative is *positive*.

The Derivative Graphically



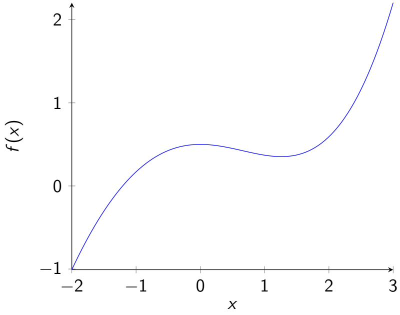
- At the region between -2 and 0, the function increases, so the derivative is *positive*.
- Between 0 and 1.3, the function decreases, so the derivative is *negative*.

The Derivative Graphically



- At the region between -2 and 0, the function increases, so the derivative is *positive*.
- Between 0 and 1.3, the function decreases, so the derivative is *negative*.
- Around 0, the function change *slowly*, so the value of the derivative is close to 0.

The Derivative Graphically



- At the region between -2 and 0, the function increases, so the derivative is *positive*.
- Between 0 and 1.3, the function decreases, so the derivative is *negative*.
- Around 0, the function change *slowly*, so the value of the derivative is close to 0.
- Around 2.5, the function changes *quickly*, so the absolute value of the derivative is high.

Definition of the derivative

- The derivative of a function is defined based on *how much the function change* if we *change the input a little bit*. This "a little bit" then becomes arbitrarily close to 0.

Definition of the derivative

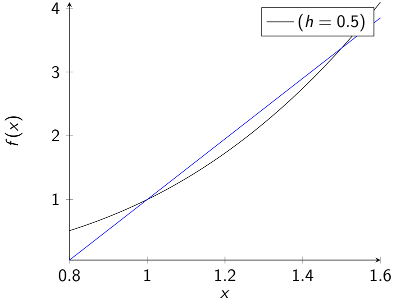
- The derivative of a function is defined based on *how much the function change* if we *change the input a little bit*. This "a little bit" then becomes arbitrarily close to 0.
- $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$

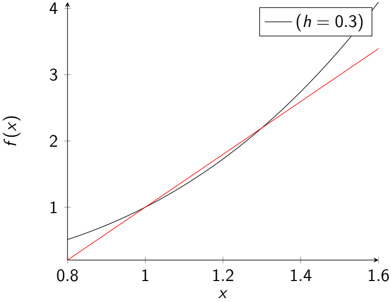
Definition of the derivative

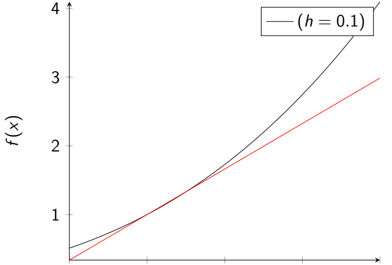
- The derivative of a function is defined based on *how much the function change* if we *change the input a little bit*. This "a little bit" then becomes arbitrarily close to 0.
- $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$
- Here h is "a little change in x, which we make go to 0 in *the limit*.

Vizulating h getting small

The lower the h the better approximation for the derivative.







Calculating the Derivative

There are many ways we can caclulate the derivative.

Calculating the Derivative

There are many ways we can caclulate the derivative.

- With *numerical* differentiation, we calculate $\frac{f(x+h)-f(x)}{h}$ for a small, but non-zero h.

Calculating the Derivative

There are many ways we can caclulate the derivative.

- With *numerical* differentiation, we calculate $\frac{f(x+h)-f(x)}{h}$ for a small, but non-zero h.
- With *exact* or *analytical* differentiation, we work with the limits mathematically and get an exact function of the derivative.

Calculating the Derivative

There are many ways we can caclulate the derivative.

- With *numerical* differentiation, we calculate $\frac{f(x+h)-f(x)}{h}$ for a small, but non-zero h.
- With *exact* or *analytical* differentiation, we work with the limits mathematically and get an exact function of the derivative.
- For example, if we work with the limits we can show that the derivative of x^3 is $3x^2$, and that $[\sin(x)]' = \cos(x)$.

Calculating the Derivative

There are many ways we can caclulate the derivative.

- With *numerical* differentiation, we calculate $\frac{f(x+h)-f(x)}{h}$ for a small, but non-zero h.
- With *exact* or *analytical* differentiation, we work with the limits mathematically and get an exact function of the derivative.
- For example, if we work with the limits we can show that the derivative of x^3 is $3x^2$, and that $[\sin(x)]' = \cos(x)$.
- We can then use these calculation as rules, apply the *chain rule*, and differentiate basically any function exectly.

Calculating the Derivative

There are many ways we can caclulate the derivative.

- With *numerical* differentiation, we calculate $\frac{f(x+h)-f(x)}{h}$ for a small, but non-zero h.
- With *exact* or *analytical* differentiation, we work with the limits mathematically and get an exact function of the derivative.
- For example, if we work with the limits we can show that the derivative of x^3 is $3x^2$, and that $[\sin(x)]' = \cos(x)$.
- We can then use these calculation as rules, apply the *chain rule*, and differentiate basically any function exectly.
- Also note that deep-learning libraries use something called *automatic differentiation*, which makes gradient for neural networks extremely efficient and easy to use. We won't cover that here, but remember to look it up if you start to use PyTorch or TensorFlow.

The Chain Rule

- Many functions can be expressed as *functions of functions*. For example, $h(x) = \sin(x^2)$ can be expressed as $g(f(x))$ where $g(x) = \sin(x)$ and $f(x) = x^2$.

The Chain Rule

- Many functions can be expressed as *functions of functions*. For example, $h(x) = \sin(x^2)$ can be expressed as $g(f(x))$ where $g(x) = \sin(x)$ and $f(x) = x^2$.
- The *chain rule* tells us how to differentiate these. It states that

$$[g(f(x))]' = g'(f(x))f'(x)$$

The Chain Rule

- Many functions can be expressed as *functions of functions*. For example, $h(x) = \sin(x^2)$ can be expressed as $g(f(x))$ where $g(x) = \sin(x)$ and $f(x) = x^2$.
- The *chain rule* tells us how to differentiate these. It states that

$$[g(f(x))]' = g'(f(x))f'(x)$$

- In our example, this means that $h'(x) = \cos(x^2) \cdot 2x$.

The Chain Rule

- Many functions can be expressed as *functions of functions*. For example, $h(x) = \sin(x^2)$ can be expressed as $g(f(x))$ where $g(x) = \sin(x)$ and $f(x) = x^2$.
- The *chain rule* tells us how to differentiate these. It states that

$$[g(f(x))]' = g'(f(x))f'(x)$$

- In our example, this means that $h'(x) = \cos(x^2) \cdot 2x$.
- This can be proved with the limits, but it can also be understood intuitively with good animations. There are no animations in this slide, so please watch *3Blue1Brown's* videos on YouTube.

The Chain Rule

- Many functions can be expressed as *functions of functions*. For example, $h(x) = \sin(x^2)$ can be expressed as $g(f(x))$ where $g(x) = \sin(x)$ and $f(x) = x^2$.
- The *chain rule* tells us how to differentiate these. It states that

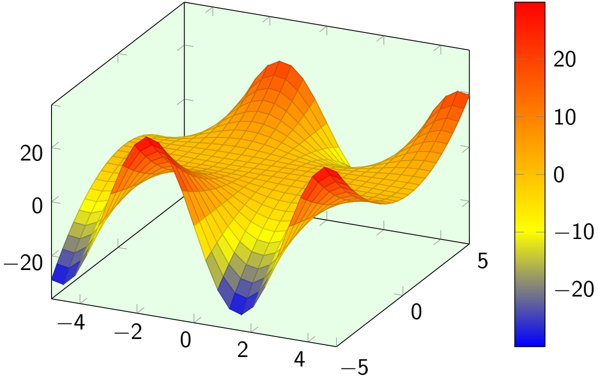
$$[g(f(x))]' = g'(f(x))f'(x)$$

- In our example, this means that $h'(x) = \cos(x^2) \cdot 2x$.
- This can be proved with the limits, but it can also be understood intuitively with good animations. There are no animations in this slide, so please watch *3Blue1Brown's* videos on YouTube.

Note: We also have rules for exponents, multiplications and many other things that we won't cover here. If you do not remember them from school, watch the videos mentioned above.

Functions of more than one variable

We now need to generalize to function of many variables.



Think of the color-value, the value upwards, as the *loss-function*, as a function of two variables (left and right), which are our *weights*.

Partial derivations

- We can now differentiate in *two* directions, the direction of w_1 and w_2.

Partial derivations

- We can now differentiate in *two* directions, the direction of w_1 and w_2.
- Our function was $f(w_1, w_2) = \sin(w_1) \cdot w_2(1 - w_2)$.

Partial derivations

- We can now differentiate in *two* directions, the direction of w_1 and w_2.
- Our function was $f(w_1, w_2) = \sin(w_1) \cdot w_2(1 - w_2)$.
- To do the derivative of with respect to w_1, we treat w_1 as our variable (x) and w_2 as a constant.

Partial derivations

- We can now differentiate in *two* directions, the direction of w_1 and w_2.
- Our function was $f(w_1, w_2) = \sin(w_1) \cdot w_2(1 - w_2)$.
- To do the derivative of with respect to w_1, we treat w_1 as our variable (x) and w_2 as a constant.
- This is the *partial derivative* of f with respect to w_1. We write $\frac{\partial f(w_1,w_2)}{\partial w_1} = \cos(w_1) \cdot w_2(1 - w_2)$.

Partial derivations

- We can now differentiate in *two* directions, the direction of w_1 and w_2.
- Our function was $f(w_1, w_2) = \sin(w_1) \cdot w_2(1 - w_2)$.
- To do the derivative of with respect to w_1, we treat w_1 as our variable (x) and w_2 as a constant.
- This is the *partial derivative* of f with respect to w_1. We write $\frac{\partial f(w_1,w_2)}{\partial w_1} = \cos(w_1) \cdot w_2(1 - w_2)$.
- Similarly, we have $\frac{\partial f(w_1,w_2)}{\partial w_2} = \sin(w_1) \cdot ((1 - w_2) - w_2)$.

The Gradient

- The gradient can be viewed as a collection of the partial derivatives.
- We have $\nabla f(w_1, w_2) = (\frac{\partial f(w_1,w_2)}{\partial w_1}, \frac{\partial f(w_1,w_2)}{\partial w_2})$.
- Note that the gradient of a function of m variable becomes an *m-dimensional vector*.
- The gradient points in the direction where the function changes the fastest.

Differentiation in Machine Learning

How do we use this in machine learning?

Differentiation in Machine Learning

How do we use this in machine learning?

- We have some *weights* that determine how we calculate the *outputs* (given some inputs).

Differentiation in Machine Learning

How do we use this in machine learning?

- We have some *weights* that determine how we calculate the *outputs* (given some inputs).
- Given some true labels (supervized learning), we can then calculate the *error* or the *loss*, which measures some difference between the output and the labels.

Differentiation in Machine Learning

How do we use this in machine learning?

- We have some *weights* that determine how we calculate the *outputs* (given some inputs).
- Given some true labels (supervized learning), we can then calculate the *error* or the *loss*, which measures some difference between the output and the labels.
- By finding the (negative) gradient to the loss function, we can then figure out where and how much we can change the weights to lower the loss function as much as possible.

Differentiation in Machine Learning

How do we use this in machine learning?

- We have some *weights* that determine how we calculate the *outputs* (given some inputs).
- Given some true labels (supervized learning), we can then calculate the *error* or the *loss*, which measures some difference between the output and the labels.
- By finding the (negative) gradient to the loss function, we can then figure out where and how much we can change the weights to lower the loss function as much as possible.
- We then use *gradient descent* (or some version of it) to iterativly update the weights).