

# Machine learning from scratch

## Lecture 2: Convex optimization

Alexis Zubiolo

`alexis.zubiolo@gmail.com`

Data Science Team Lead @ Adcash

February 2, 2017

# Optimization: Derivatives

The **derivative** is an important concept in optimization and in machine learning. Mathematical definition: Let  $f : \mathbb{R} \mapsto \mathbb{R}$  be a function. Its derivative  $f'$  is defined by:

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

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

**Case 1:**  $f'(x) > 0$

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

**Case 1:**  $f'(x) > 0$

Suppose  $h > 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is increasing (because } x+h > x)\end{aligned}$$

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

**Case 1:**  $f'(x) > 0$

Suppose  $h > 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is increasing (because } x+h > x)\end{aligned}$$

Now, suppose  $h < 0$ , we have

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

**Case 1:**  $f'(x) > 0$

Suppose  $h > 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is increasing (because } x+h > x\text{)}\end{aligned}$$

Now, suppose  $h < 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is increasing (because } x+h < x\text{)}\end{aligned}$$

## Derivatives: Why we use them (1)

**Derivatives** give us an idea of the local behavior of the function.

**Case 1:**  $f'(x) > 0$

Suppose  $h > 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is increasing (because } x+h > x\text{)}\end{aligned}$$

Now, suppose  $h < 0$ , we have

$$\begin{aligned}f'(x) > 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} > 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is increasing (because } x+h < x\text{)}\end{aligned}$$

**Conclusion:**  $f$  is increasing  $\iff f' > 0$



## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

Suppose again  $h < 0$ , we have:

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

Suppose again  $h < 0$ , we have:

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

## Derivatives: Why we use them (2)

**Case 2:**  $f'(x) < 0$

Now, suppose again  $h > 0$ , we have

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) < f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

Suppose again  $h < 0$ , we have:

$$\begin{aligned}f'(x) < 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} < 0 \\&\iff f(x+h) > f(x) \\&\iff f \text{ is decreasing}\end{aligned}$$

**Conclusion:**  $f$  is decreasing  $\iff f' < 0$

## Derivatives: Why we use them (3)

**Case 3:**  $f'(x) = 0$ , then



## Derivatives: Why we use them (3)

**Case 3:**  $f'(x) = 0$ , then

$$\begin{aligned}f'(x) = 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = 0 \\&\iff f(x+h) = f(x) \\&\iff f \text{ is constant}\end{aligned}$$

## Derivatives: Why we use them (3)

**Case 3:**  $f'(x) = 0$ , then

$$\begin{aligned}f'(x) = 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = 0 \\&\iff f(x+h) = f(x) \\&\iff f \text{ is constant}\end{aligned}$$

**Conclusion:**  $f$  is constant  $\iff f' = 0$

## Derivatives: Why we use them (3)

**Case 3:**  $f'(x) = 0$ , then

$$\begin{aligned}f'(x) = 0 &\iff \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = 0 \\&\iff f(x+h) = f(x) \\&\iff f \text{ is constant}\end{aligned}$$

**Conclusion:**  $f$  is constant  $\iff f' = 0$

**Summary:**

- ▶  $f$  is increasing  $\iff f'(x) > 0$
- ▶  $f$  is decreasing  $\iff f'(x) < 0$
- ▶  $f$  is constant  $\iff f'(x) = 0$

## Derivatives: Why we use them (4)

### Summary:

- ▶  $f$  is increasing  $\iff f'(x) > 0$
- ▶  $f$  is decreasing  $\iff f'(x) < 0$
- ▶  $f$  is constant  $\iff f'(x) = 0$

## Derivatives: Why we use them (4)

### Summary:

- ▶  $f$  is increasing  $\iff f'(x) > 0$
- ▶  $f$  is decreasing  $\iff f'(x) < 0$
- ▶  $f$  is constant  $\iff f'(x) = 0$

What do we do with this?

# Derivatives: Why we use them (4)

## Summary:

- ▶  $f$  is increasing  $\iff f'(x) > 0$
- ▶  $f$  is decreasing  $\iff f'(x) < 0$
- ▶  $f$  is constant  $\iff f'(x) = 0$

What do we do with this?

If we want to find  $f$ 's minimum(s), a strategy could be:

1. Start at a random point  $x_0 \in \mathbb{R}$
2. Compute  $f'(x_0)$  and then
  - ▶ if  $f'(x_0) > 0$  then move to the left (because  $f$  is increasing)
  - ▶ if  $f'(x_0) < 0$  then move to the right (because  $f$  is decreasing)
  - ▶ if  $f'(x_0) = 0$  then we stop

## Derivatives: How to compute them?

There are several ways to compute the derivatives.

# Derivatives: How to compute them?

There are several ways to compute the derivatives.

- By **applying the definition**

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

with a small  $h$  (e.g.  $h = 0.01$ ). This is called the **finite difference approximation**.

- By using **closed-form derivatives**, e.g.

$$\text{if } f(x) = x^2, \text{ then } f'(x) = 2x$$

How do we know that? By using the formula!



## Derivative: Practical example

Recall the definition:

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

**Exercise:** Apply this formula for  $f(x) = x^2$ . What is  $f'(x)$ ?

## Derivative: Practical example

Recall the definition:

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

**Exercise:** Apply this formula for  $f(x) = x^2$ . What is  $f'(x)$ ?

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

We can apply this logic to the usual functions (log, cos, sin, ...) and obtain what we call **closed-form solutions**.

## Derivative: Practical example

Recall the definition:

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

**Exercise:** Apply this formula for  $f(x) = x^2$ . What is  $f'(x)$ ?

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

We can apply this logic to the usual functions (log, cos, sin, ...) and obtain what we call **closed-form solutions**.

**Exercise:** Do the same for  $f(x) = \frac{1}{x}$ .

## Derivatives: Finite difference approximation

**Recall:** The **finite difference approximation** consists in applying the following formula

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

with  $h$  close enough to 0.

## Derivatives: Finite difference approximation

**Recall:** The **finite difference approximation** consists in applying the following formula

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

with  $h$  close enough to 0.

There's a symmetric and more stable formula, called the **centered finite difference approximation**.

$$f'(x) = \frac{df}{dx}(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x-h)}{2h}$$

## Chain rule

Another important rule concerning derivatives is the **chain rule**:

$$(f(g(x)))' = g'(x)f'(g(x)) \quad (1)$$

or, equivalently:

$$\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \frac{dg(x)}{dx} \quad (2)$$

## Gradient: Generalizing the derivatives

Derivatives look like a practical way to get the minimum of a function, which is what we want to achieve (minimizing  $J(\theta)$ ).

## Gradient: Generalizing the derivatives

Derivatives look like a practical way to get the minimum of a function, which is what we want to achieve (minimizing  $J(\theta)$ ).

However, the definition we've seen only applies to function from  $\mathbb{R}$  to  $\mathbb{R}$ . The **gradient** is a generalization of the derivative for functions from  $\mathbb{R}^d$  to  $\mathbb{R}$ , like  $J$ .



## Beyond derivatives: Gradient

Let's consider a function of  $f$  2 real variables  $x$  and  $y$ , e.g.

$$f(x, y) = xy$$

The gradient of  $f$  is a 2-dimensional vector noted  $\nabla f$  defined by:

$$\nabla f(x, y) = \left[ \frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y} \right]$$

where

$$\frac{\partial f(x, y)}{\partial x}$$

is the **partial derivative** of  $f$  with respect to  $x$  (considering  $y$  as a constant), and

$$\frac{\partial f(x, y)}{\partial y}$$

is the **partial derivative** of  $f$  with respect to  $y$  (considering  $x$  as a constant).

## Beyond derivatives: Gradient

Let's consider a function of  $f$  2 real variables  $x$  and  $y$ , e.g.

$$f(x, y) = xy$$

The gradient of  $f$  is a 2-dimensional vector noted  $\nabla f$  defined by:

$$\nabla f(x, y) = \left[ \frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y} \right]$$

where

$$\frac{\partial f(x, y)}{\partial x}$$

is the **partial derivative** of  $f$  with respect to  $x$  (considering  $y$  as a constant), and

$$\frac{\partial f(x, y)}{\partial y}$$

is the **partial derivative** of  $f$  with respect to  $y$  (considering  $x$  as a constant).

**Exercise:** What is  $\nabla f(x, y)$  for  $f(x, y) = xy$  and  $f(x, y) = x + y$ ?

## Gradient: Generalizing the derivatives

Recall we want to minimize the following cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left( h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2$$

## Gradient: Generalizing the derivatives

Recall we want to minimize the following cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left( h \left( \mathbf{x}^{(i)} \right) - y^{(i)} \right)^2$$

To do so, we need to compute  $\nabla J(\theta)$ , the **gradient** of  $J$  in  $\theta$ , which is, by definition:

$$\nabla J(\theta) = \left[ \frac{\partial}{\partial \theta_1} J(\theta), \dots, \frac{\partial}{\partial \theta_d} J(\theta) \right]^T$$

## Gradient: Generalizing the derivatives

Recall we want to minimize the following cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left( h \left( \mathbf{x}^{(i)} \right) - y^{(i)} \right)^2$$

To do so, we need to compute  $\nabla J(\theta)$ , the **gradient** of  $J$  in  $\theta$ , which is, by definition:

$$\nabla J(\theta) = \left[ \frac{\partial}{\partial \theta_1} J(\theta), \dots, \frac{\partial}{\partial \theta_d} J(\theta) \right]^T$$

Interpretation of the gradient: Direction of **steepest descent** of  $f$

## Back to least squares: context reminder

living area (m <sup>2</sup> )	# bedrooms	intercept	price (1000's BGN)
50	1	1	30
76	2	1	48
26	1	1	12
102	3	1	90

$$h(\mathbf{x}) = \sum_{j=0}^d \theta_j x_j = \boldsymbol{\theta}^T \mathbf{x}$$

## Back to least squares: context reminder

living area (m <sup>2</sup> )	# bedrooms	intercept	price (1000's BGN)
50	1	1	30
76	2	1	48
26	1	1	12
102	3	1	90

$$h(\mathbf{x}) = \sum_{j=0}^d \theta_j x_j = \boldsymbol{\theta}^T \mathbf{x}$$

Suppose we chose the following loss function:

$$\ell(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

This leads to the following least squares *cost function*:

$$J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n \left( h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2$$

This problem the **ordinary least squares** (OLS) regression model.

## Least Mean Squares (LMS) update rule

To apply the LMS update rule, we need to compute the gradient of  $J$ . Let's compute it for a single  $(\mathbf{x}, y)$  sample:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h(\mathbf{x}) - y)^2 \\ &= ?\end{aligned}$$

where  $h(\mathbf{x}) = \sum_{j=0}^d \theta_j x_j = \theta^T \mathbf{x}$



## Least Mean Squares (LMS) update rule

To apply the LMS update rule, we need to compute the gradient of  $J$ . Let's compute it for a single  $(\mathbf{x}, y)$  sample:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h(\mathbf{x}) - y)^2 \\ &= ?\end{aligned}$$

where  $h(\mathbf{x}) = \sum_{j=0}^d \theta_j x_j = \theta^T \mathbf{x}$

**Exercise:** Compute the gradient and find the update rule.

## Least Mean Squares (LMS) update rule

**Solution:**

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h(\mathbf{x}) - y)^2 \\&= 2 \frac{1}{2} (h(\mathbf{x}) - y) \frac{\partial}{\partial \theta_j} (h(\mathbf{x}) - y) \\&= (h(\mathbf{x}) - y) \frac{\partial}{\partial \theta_j} \left( \sum_{k=0}^d \theta_k x_k - y \right) \\&= (h(\mathbf{x}) - y) x_j\end{aligned}$$

## Least Mean Squares (LMS) update rule

**Solution:**

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h(\mathbf{x}) - y)^2 \\&= 2 \frac{1}{2} (h(\mathbf{x}) - y) \frac{\partial}{\partial \theta_j} (h(\mathbf{x}) - y) \\&= (h(\mathbf{x}) - y) \frac{\partial}{\partial \theta_j} \left( \sum_{k=0}^d \theta_k x_k - y \right) \\&= (h(\mathbf{x}) - y) x_j\end{aligned}$$

So the gradient descent update becomes

$$\begin{aligned}\theta_j &:= \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\&:= \theta_j + \alpha (y - h(\mathbf{x})) x_j\end{aligned}$$

# Conclusion

We had an overview of the optimization techniques.

# Conclusion

We had an overview of the optimization techniques.

Next Thursday, we will review what needs to be reviewed (tell me by email so that I can adapt or when the class starts) and start implementing:

- ▶ The loss function for the least-squares problem
- ▶ Its gradient
- ▶ The Gradient descent weight upgrade
- ▶ Run it on a toy example

Thank you! Questions?

`alexis.zubiollo@gmail.com`

`https://github.com/azubiollo/itstep`