

# 06-20416 and 06-12412 (Intro to) Neural Computation

## 04 – Gradient Descent

**Per Kristian Lehre**

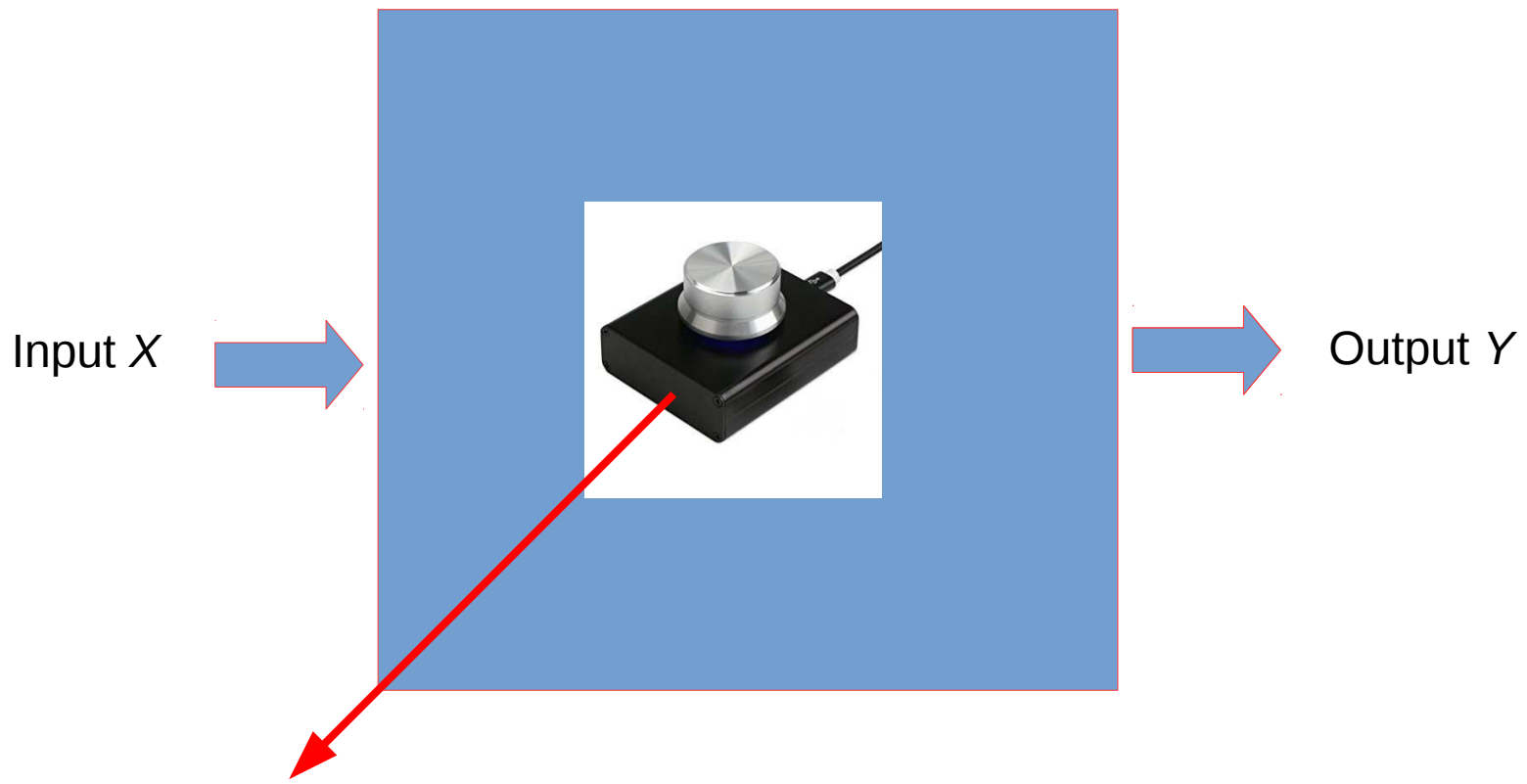
# Last lecture

---

- Probabilistic models
- Some probabilistic concepts
  - Random variable, density function, normal distribution, joint density function, empirical distribution
- Maximum likelihood
  - Likelihood function and Maximum likelihood estimate
  - Learning via log-likelihood. Example: linear regression

# So far: Adjusting a single knob

---

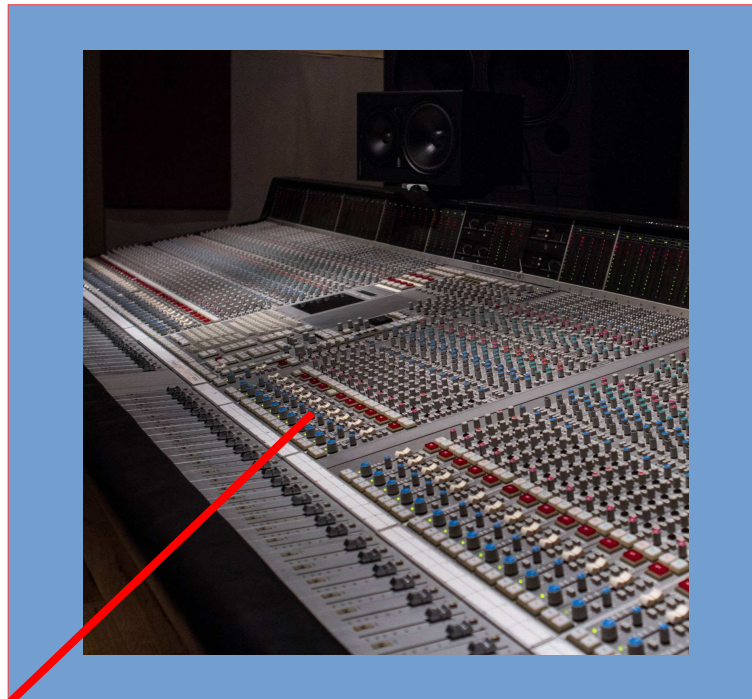
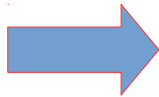


$$J(\Theta) = \mathbb{E}_{(X,Y) \sim \hat{\mathcal{D}}} - \log p_{\text{model}}(Y \mid X; \Theta)$$

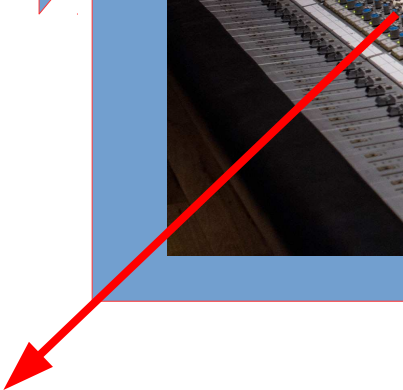
# Today: Adjusting millions of knobs

---

Input



Output


$$J(\Theta) = \mathbb{E}_{(X,Y) \sim \hat{\mathcal{D}}} - \log p_{\text{model}}(Y \mid X; \Theta)$$

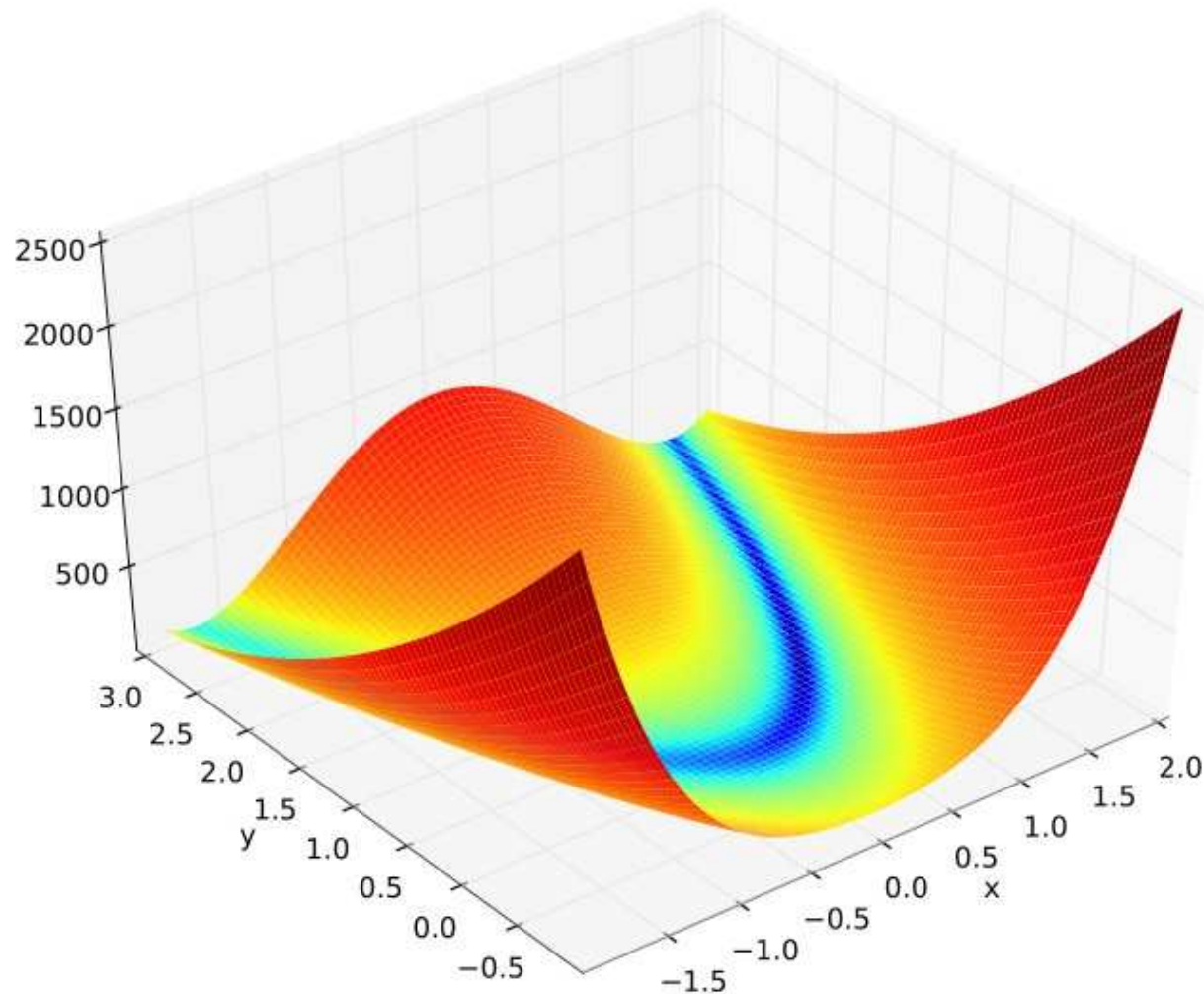
# Outline

---

- Functions of multiple variables
- Partial derivatives and the chain rule
- Gradients
  - Direction of steepest ascent
- Gradient descent

# Functions of Multiple Variables

---

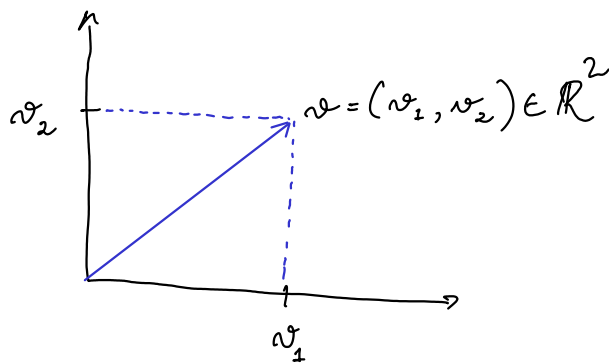


## Vectors

Vectors are "arrays" of numbers, e.g.

$$v = (v_1, \dots, v_m) \in \mathbb{R}^m$$

We can consider a vector as a point in space, where each element  $v_i$  giving the coordinate along the  $i$ -th axis, e.g.



Norms assigns "lengths" to vectors.

The  $L^p$ -norm of a vector  $v \in \mathbb{R}^m$  is

$$\|v\|_p = \left( \sum_i |v_i|^p \right)^{1/p}$$

with  $p \in \mathbb{R}$  and  $p \geq 1$ .

The special case  $L^2$  is the Euclidean norm, denoted  $\|v\| = \|v\|_2$ .

## Operations on vectors

For all  $a \in \mathbb{R}$ ,  $u = (u_1, \dots, u_m) \in \mathbb{R}^m$ , and  $v = (v_1, \dots, v_m) \in \mathbb{R}^m$ ,

$$au = (au_1, \dots, au_m)$$

scalar multiplication

$$u + v = (u_1 + v_1, \dots, u_m + v_m)$$

vector addition

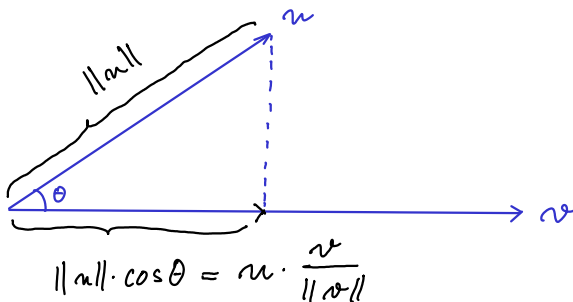
$$u \cdot v = \sum_{i=1}^m u_i v_i$$

dot product

## Theorem (Geometric Interpretation of Dot Product)

If the angle between two vectors  $u, v \in \mathbb{R}^m$  is  $\theta$ , then

$$u \cdot v = \|u\| \cdot \|v\| \cdot \cos \theta$$





## Partial Derivative

The partial derivative of a function

$$f(x_1, \dots, x_m)$$

in the direction of variable  $x_i$  at the point  $a = (a_1, \dots, a_m)$  is

$$\frac{\partial f}{\partial x_i}(a_1, \dots, a_m) = \lim_{h \rightarrow 0} \frac{f(a_1, \dots, a_i + h, \dots, a_m) - f(a_1, \dots, a_m)}{h}$$

Intuitively the derivative of a function  $g(x_i) = f(x_1, \dots, x_n)$ , where all variables except  $x_i$  are fixed as constants.

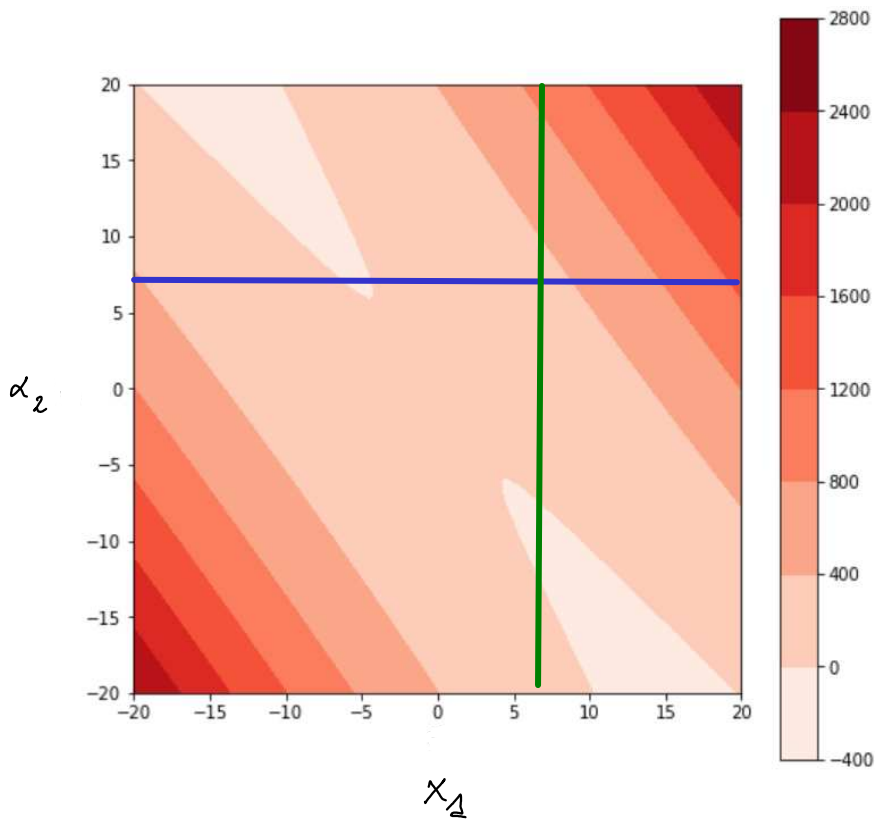
### Example

$$f(x_1, x_2) = 2x_1^2 + x_2^2 + 3x_1x_2 + 4$$

$$\frac{\partial f}{\partial x_1} = 4x_1 + 3x_2$$

$$\frac{\partial f}{\partial x_2} = 2x_2 + 3x_1$$

# Partial Derivatives : Geometric Interpretation



$\frac{\partial f}{\partial x_1}$  is the rate of change of  $f$  along dimension  $x_1$  (i.e., blue line)

$\frac{\partial f}{\partial x_2}$  is the rate of change of  $f$  along dimension  $x_2$  (i.e., green line)

# Gradient

## Definition

The gradient of a function  $f(x_1, \dots, x_m)$  is

$$\nabla f := \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_m} \right)$$

## Example

$$f(x_1, x_2) = 2x_1^2 + x_2^2 + 3x_1x_2 + 4$$

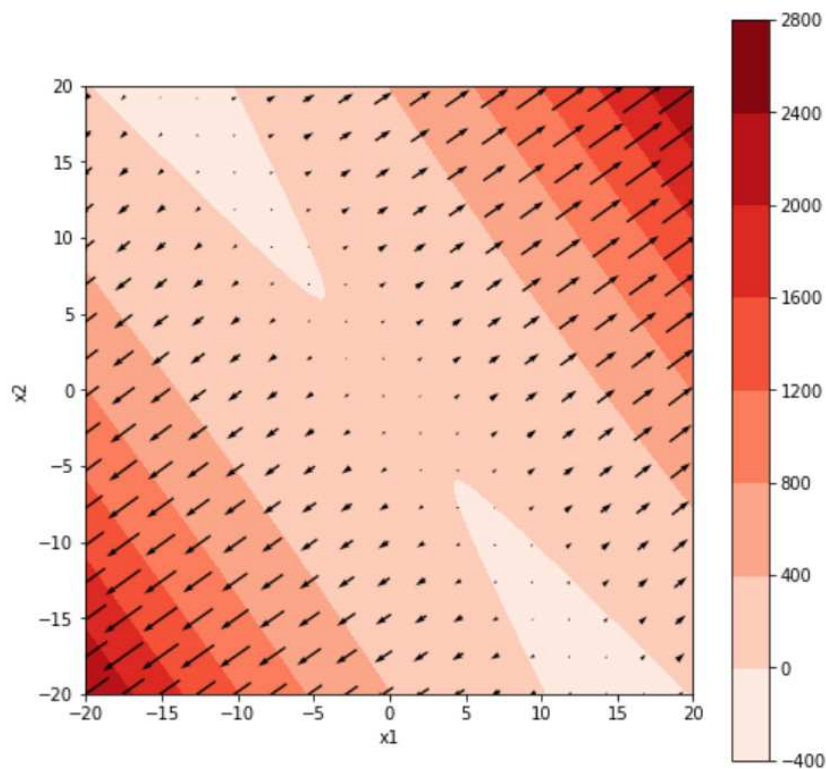
$$\nabla f(x_1, x_2) = \left( \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right)$$

$$= (4x_1 + 3x_2, 2x_2 + 3x_1)$$

# Visualisation of the Gradient

Remark

If  $f: \mathbb{R}^m \rightarrow \mathbb{R}$ , then  $\nabla f: \mathbb{R}^m \rightarrow \mathbb{R}^m$ ,  
i.e., the gradient is a vector-valued  
function that maps vectors to vectors



## Chain Rule (special case)

For one-dimensional functions

If  $y = f(u)$  and

$$u = g(x)$$

then

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

For higher dimensional functions

If  $y = f(u_1, \dots, u_m)$  and

$$u_i = g_i(x_1, \dots, x_n) \quad \text{for } i \in \{1, \dots, m\}$$

then

$$\frac{\partial y}{\partial x_i} = \sum_{j=1}^m \frac{\partial y}{\partial u_j} \cdot \frac{\partial u_j}{\partial x_i}$$

## Chain Rule (Example)

$$h(x_1, x_2) = (ax_1 + bx_2)^2 x_1 x_2$$

We can express  $h$  by defining

$$y = f(u_1, u_2) := u_1^2 \cdot u_2$$

$$u_1 = g_1(x_1, x_2) := ax_1 + bx_2$$

$$u_2 = g_2(x_1, x_2) := x_1 x_2$$

Applying the chain rule gives

$$\begin{aligned}\frac{\partial h}{\partial x_1} &= \frac{\partial y}{\partial x_1} = \frac{\partial f}{\partial u_1} \cdot \frac{\partial u_1}{\partial x_1} + \frac{\partial f}{\partial u_2} \cdot \frac{\partial u_2}{\partial x_1} \\&= 2u_1 u_2 \cdot a + u_1^2 \cdot x_2 \\&= u_1 (2au_2 + u_1 x_2) \\&= (ax_1 + bx_2)(2a x_1 x_2 + (ax_1 + bx_2)x_2) \\&= (ax_1 + bx_2)x_2(3ax_1 + bx_2)\end{aligned}$$

Similarly, the chain rule gives

$$\begin{aligned}\frac{\partial h}{\partial x_2} &= \frac{\partial y}{\partial x_2} = \frac{\partial f}{\partial u_1} \cdot \frac{\partial u_1}{\partial x_2} + \frac{\partial f}{\partial u_2} \cdot \frac{\partial u_2}{\partial x_2} \\&= 2u_1 u_2 b + u_1^2 x_1 \\&= u_1 (2bu_2 + u_1 x_1) \\&= (ax_1 + bx_2)(2bx_1 x_2 + (ax_1 + bx_2)x_1) \\&= (ax_1 + bx_2)x_1(3bx_2 + ax_1)\end{aligned}$$

# Directional Derivative

## Definition

Given a function

$$f: \mathbb{R}^m \rightarrow \mathbb{R}$$

and a vector

$$v = (v_1, \dots, v_m), \text{ with } \|v\| = 1$$

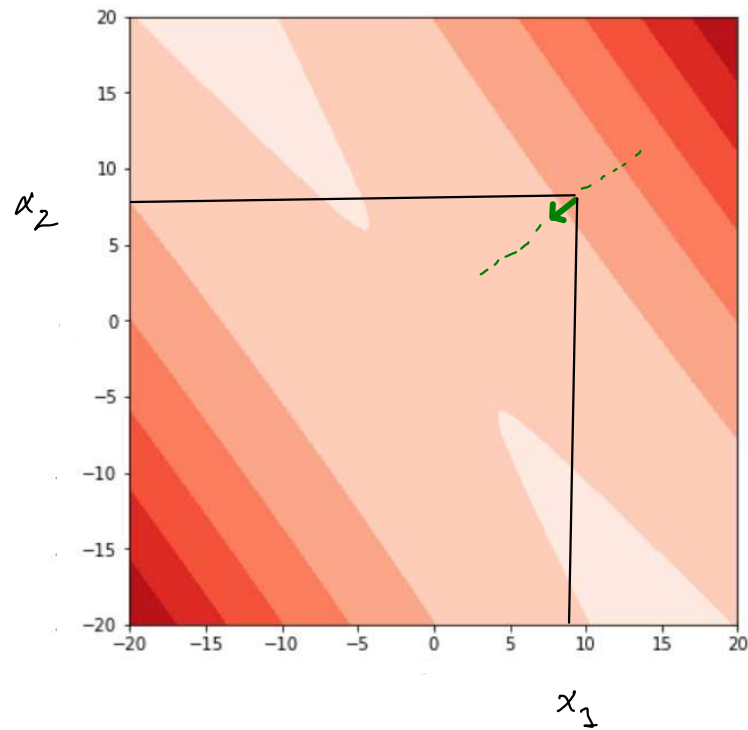
the directional derivative of  $f$  in

$$x = (x_1, \dots, x_m)$$

along the vector  $v$  is

$$\nabla_v f(x) := \lim_{\alpha \rightarrow 0} \frac{f(x + \alpha v) - f(x)}{\alpha}$$

$$= \lim_{\alpha \rightarrow 0} \frac{f(x_1 + \alpha v_1, \dots, x_m + \alpha v_m) - f(x_1, \dots, x_m)}{\alpha}$$



## Computing Directional Derivative

The following theorem implies that if we know the gradient  $\nabla f$ , then we can compute the derivative in any direction  $v$ .

### Theorem

$$\nabla_v f(x) = \nabla f(x) \cdot v$$

directional derivative

dot product

gradient

Proof

Define the function

$$h(\alpha) := f(u_1, \dots, u_m)$$

where

$$u_i := x_i + \alpha v_i \quad \text{for all } i \in \{1, \dots, m\}.$$

Note that  $h: \mathbb{R} \rightarrow \mathbb{R}$ , i.e.,  $h$  is a one-dimensional real-valued function.

$$\begin{aligned} \nabla_v f(x) &= \lim_{\alpha \rightarrow 0} \frac{f(x + \alpha v) - f(x)}{\alpha} \\ &= \lim_{\alpha \rightarrow 0} \frac{h(0 + \alpha) - h(0)}{\alpha} \end{aligned}$$

by def. of  $\nabla_v f$

by def. of  $g$

by def. of derivative

$$(1) \quad = h'(0)$$

Using the chain rule, we have

$$(2) \quad h'(\alpha) = \frac{dh}{d\alpha} = \sum_{i=1}^m \frac{\partial f}{\partial u_i} \cdot \frac{\partial u_i}{\partial \alpha} = \sum_{i=1}^m \frac{\partial f}{\partial u_i} \cdot v_i$$

Note that for  $\alpha=0$ , we have

$$(3) \quad u_i = x_i + 0 \cdot v_i = x_i$$

Using (1), (2), and (3), we get

$$\nabla_v f(x) = h'(0) = \sum_{i=1}^m \frac{\partial f}{\partial x_i} \cdot v_i = \nabla f(x) \cdot v$$





# The Gradient Points towards Steepest Ascent

The vector  $v$  along which  $f$  has steepest ascent is

$$\operatorname{argmax}_{v, \|v\|=1} \nabla_v f(x)$$

$$= \operatorname{argmax}_{v, \|v\|=1} \nabla f(x) \cdot v$$

angle between  
 $v$  and  $\nabla f(x)$

$$= \operatorname{argmax}_{v, \|v\|=1} \|\nabla f(x)\| \|v\| \cos \theta$$

$$= \operatorname{argmax}_{v, \|v\|=1} \|\nabla f(x)\| \cos \theta$$

$\Rightarrow$  The vector  $v$  which gives the steepest ascent is the vector that has angle  $\theta=0$  to  $\nabla f$ , i.e., the vector  $v$  which points in the same direction as  $\nabla f$ .

# Method of Gradient Descent

NB! This is the most important slide in this module!

Input: cost function  $J: \mathbb{R}^m \rightarrow \mathbb{R}$   
learning rate  $\epsilon \in \mathbb{R}, \epsilon > 0$

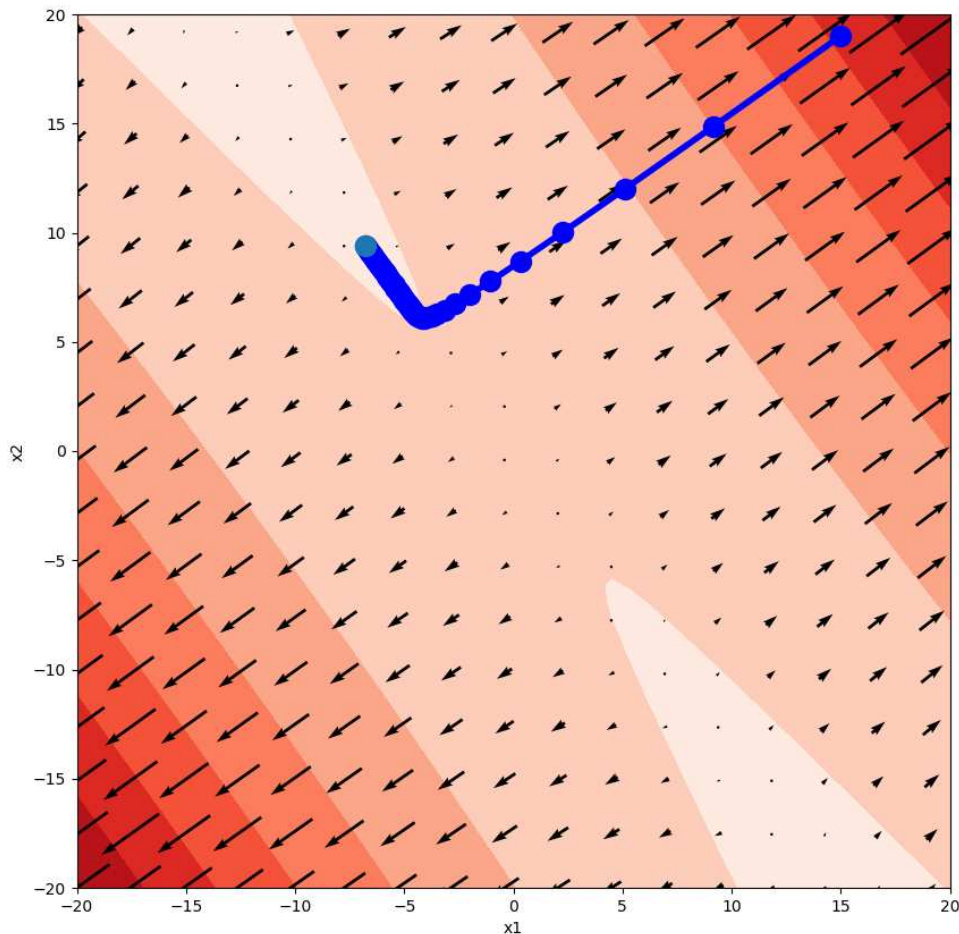
$x \leftarrow$  some initial point in  $\mathbb{R}^m$   
while termination condition not met {

$$x \leftarrow x - \epsilon \cdot \nabla J(x)$$

}



A. Cauchy (1789-1857)



# Summary

---

- Functions of multiple variables
- Partial derivatives and the chain rule
- Gradients
  - Direction of steepest ascent
- Gradient descent

# Next time

---

- Feed forward neural networks
- The backpropagation algorithm