4. Optimization for functions of several variables I: Leas Squares & Machine Learning

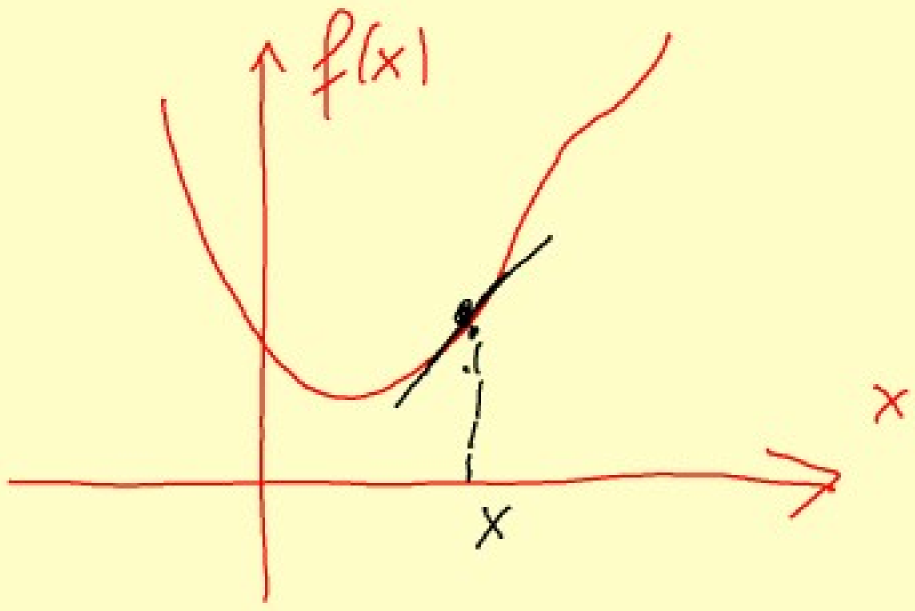
The Fréchet differential is a linear function T s.t.
of $f : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$

$$\lim_{y \to x} \frac{|f(y) - f(x) - T(x-y)|}{\| y - x \|} = 0$$

Meaning of Fréchet diff: you approx. f by a lin. function (locally in x!)

(Notation $df(x)(z) \overset{\text{Def}}{=} T(z)$

linear function in the limit above



T1. If all $\frac{\partial}{\partial x_i} f$ are continuous at x then f is Fréchet diffable at x and

$$df(x)(z) = \nabla f(x) \cdot z \qquad \forall z \in \mathbb{R}^n$$

Furthermore if all $\frac{\partial^2}{\partial x_i \partial x_j} f$ are cont. at x then 2^{nd} Fréchet diff is a quadratic function (form) with matrix $a_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$ (Hesse of f).

Analogy

$d = 1$

$f: \mathbb{R} \to \mathbb{R}$

f'.

f''

$d > 1$ (high dim)

$f: \mathbb{R}^d \to \mathbb{R}$ several vars.

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

$$H_f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j=\overline{1,d}}$$

Notation $\parallel$

$$\nabla^2 f$$

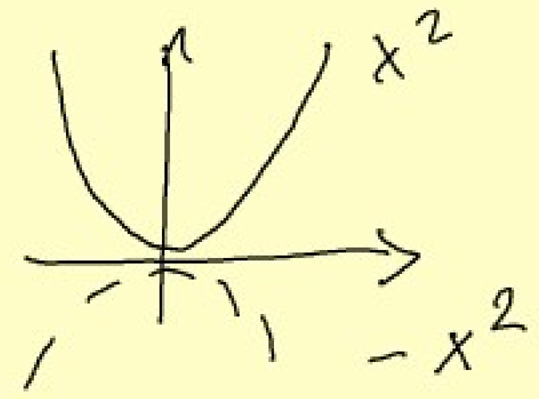
§4.1. Optimization for functions of several variables

For functions of a single var ($d=1$)

If $f'(x^*) = 0$ and $f''(x^*) > 0 \Rightarrow$ x^* min (local)

$f'(x^*) = 0$ and $f''(x^*) < 0 \Rightarrow$ x^* max

How to remember this: think about the simplest function(s) $f(x) = x^2$ (or $f(x) = -x^2$)



$f''(x) = 2 > 0$ (or $f''(x) = -2 < 0$)

Th 2. (FERMAT)

$f: \mathbb{R}^d \to \mathbb{R}$, F diff-able in $x^* \in \mathbb{R}^d$.

If x^* is a local min/max then $\nabla f(x^*) = 0$.

(F-diff-able = Fréchet - ...)

(Optimization = find minima/maxima)

The FERMAT classical approach to optimiz:

compute f', find x^* s.t. $f'(x^*) = 0$

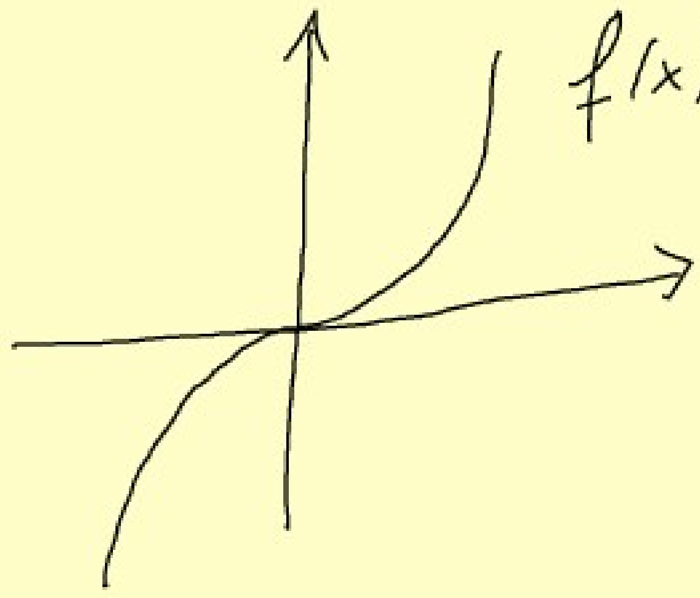
Then, compute (establish) sign of $f''(x^*) \overset{?}{\gtrless} 0$

Rk.

There exist critical points ($f' = 0$ or $\nabla f = 0$) which are neither minima nor maxima.

$d = 1$

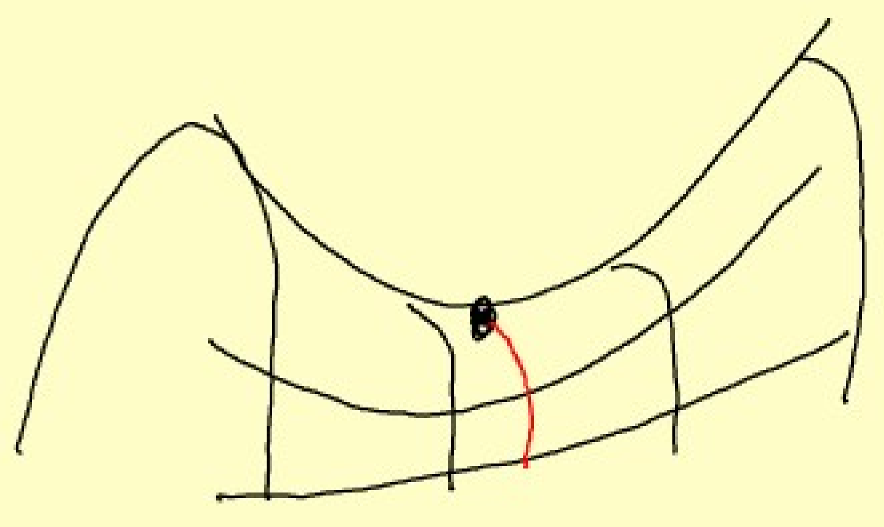
$f(x) = x^3$



$d = 2$

$f(x_1, x_2) = x_1^2 - x_2^2$

($(0,0)$ saddle point)



Positivity for functions of several variables?

Def: A quadratic function (form) $Q : \mathbb{R}^n \to \mathbb{R}$ (with matrix $A = (a_{ij})$)

is positive definite if $Q(x) > 0 \quad \forall x \in \mathbb{R}^d \setminus \{0_{\mathbb{R}^d}\}$

negative def $\quad Q(x) < 0 \quad$ — // —

indefinite if $Q(x_1) > 0, \; Q(x_2) < 0$

also we say that Q is

positive <u>semi</u> def if $\quad Q(x) \geq 0$

negative <u>semi</u> def $\quad Q(x) \leq 0$

Th 3 (SYLVESTER) Crit. for posi/neg def

if $A = (a_{ij})$ is the matrix of Q

Then

- $a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \ldots, \begin{vmatrix} a_{11} & \cdots & \\ \vdots & \ddots & \\ & & a_{dd} \end{vmatrix} > 0$

  $\Rightarrow$ Q is posi def.

- $a_{11} < 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0 \quad \ldots \quad (-1)^d \begin{vmatrix} a_{11} & \cdots & \\ \vdots & \ddots & \\ & & a_{dd} \end{vmatrix} > 0$

  $\Rightarrow$ Q is neg def

  (signs alternate)

- otherwise criterion is not effective.

T4. $f: \mathbb{R}^d \to \mathbb{R}$ twice F-diff-able in x^*

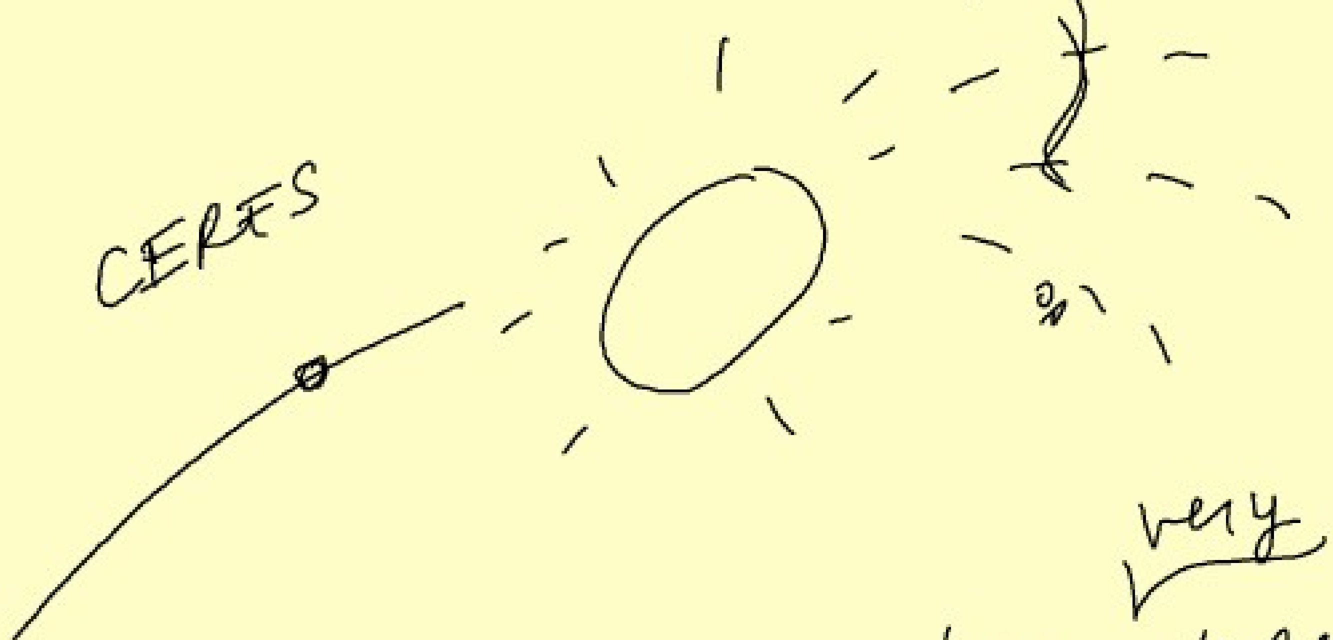
if $\nabla f(x^*) = 0_{\mathbb{R}^d}$ and

$H_f(x^*) = \nabla^2 f(x^*)$ is pos def $\Rightarrow x^*$ min
neg def $\Rightarrow x^*$ max.

§4.2. The Least Squares Method (GAUSS)

1801 Piazzi was following the asteroid CERES

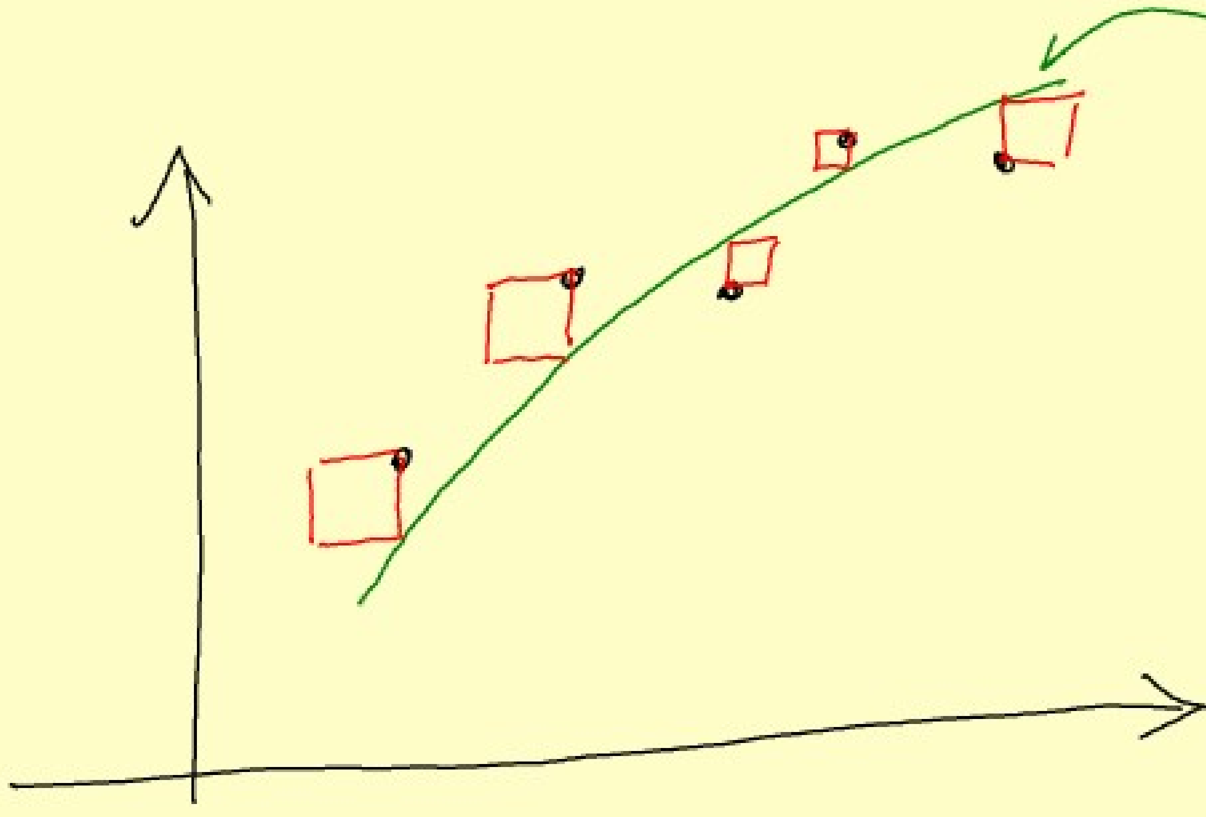
which dissap. behind the sun



Question: where will we find it

GAUSS makes very precise prediction by using the Least Squares Method.

Idea



"closest" to the whole data set

minimize area of squares ("squared errors")

Formal Statement

Given • a set of data (measurement)

x	x_1	x_i	x_m
y	$y_1 \ldots$	$y_i \ldots$	y_m

• a model $f(x) = a x + b$

= a parametrized family of functions

Goal: Find a^*, b^* such that $a^* x + b^*$ is the best fit for the given data.

This is an Optimization Problem!

$$E(a,b) = \sum_{i=1}^{m} \left(y_i - (a x_i + b) \right)^2 \to \min$$

("Least squares")

minimize w.r.t. a, b!

Rk. E is "quadratic" $\Rightarrow$ $\nabla^2 E$ pos. def

So you only have to find a^*, b^* such that

$$\nabla E(a^*, b^*) = 0 \iff \begin{cases} \dfrac{\partial E}{\partial a}(a^*, b^*) = 0 \\ \dfrac{\partial E}{\partial b}(a^*, b^*) = 0 \end{cases}$$

[D. Popa, p. 190-1]

$$\begin{cases} \sum_{i=1}^{m} (y_i - a x_i - b)\, x_i = 0 \\ \sum_{i=1}^{m} (y_i - a x_i - b) = 0 \end{cases}$$

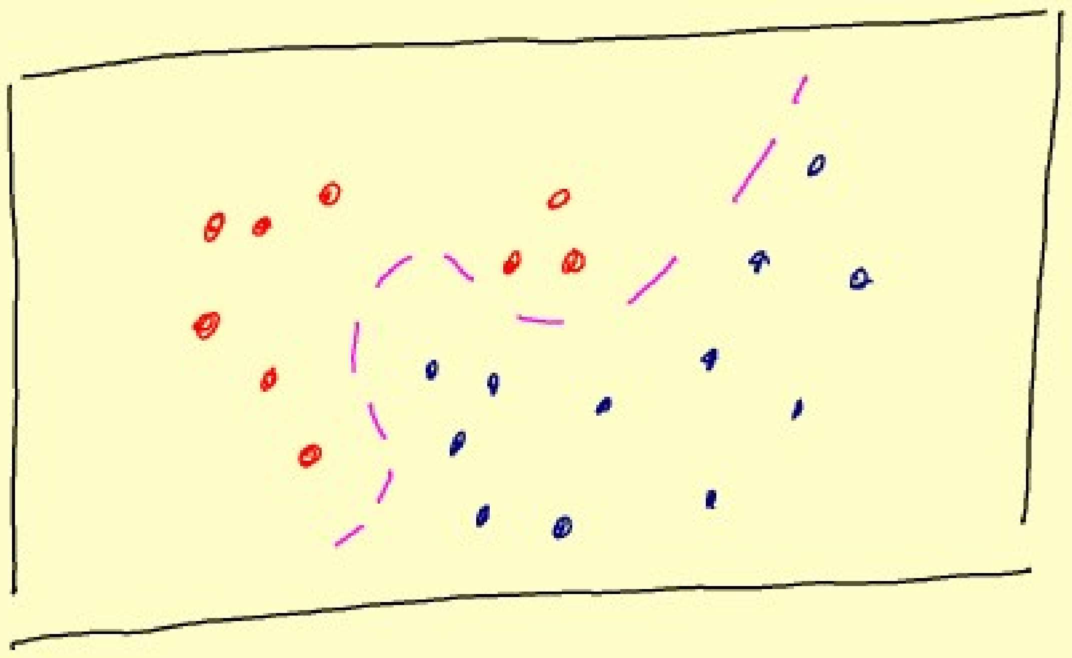
2×2 lin. system (explicitly solvable)

$f = a^* x + b^*$ Regression line

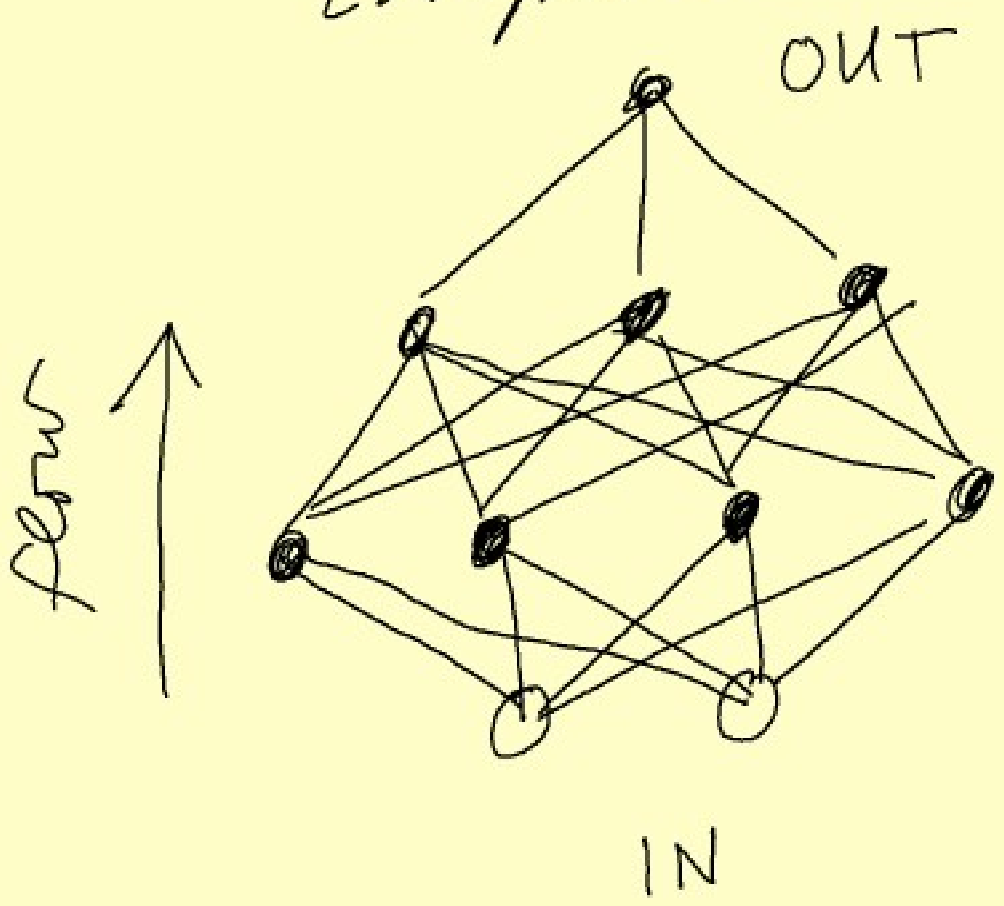
§4.3: Deep Learning

Example of Classification Problem

Want: train computer to "learn" frontier betw. blue & red



computer = artificial Neural Network



Out put Layer (1 neuron in my case)

Hidden Layers (active neurons)

input Layer

each active neuron has an activation function (SIGMOID)

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

(x = input of neuron)
($\phi(x)$ = output)

connected neurons

$$\phi\Big(\sum_j \underbrace{w_{ij}}_{\text{weights}} \underbrace{x_j}_{} + \underbrace{b_i}_{\text{bias}}\Big) = \text{output of neuron "}i\text{"}$$

input from neurons on previous layer

$y_{OUT} = F(x_{IN})$

entire NN = function connecting IN to OUT

—

in our example IN : $x_{IN} = (x_{1_{IN}}, x_{2_{IN}})$

coordinates of a point on the map

"Classification" OUT y_{OUT} = number $\in [0,1]$

0 = you are in the red zone

1 = — " — blue zone

<u>Training</u> the NN given the labeled dataset

$(x^i, y^i) \quad i = 1, m$

points

labels $\in \{0,1\}$

IN to OUT function of NN

Apply Least Squares to

$$E(\underline{w}, \underline{b}) = \sum_{i=1}^{m} \left(y^i - F_{\underline{w},\underline{b}}(x^i) \right)^2 \rightarrow \min$$

parameters = weights w and biases b

HOW to minimize E ?

GRADIENT DESCENT ! (Algorithm)

for $W = (\underline{w}, \underline{b})$

"Learning rate"

(GD) $$W_{n+1} = W_n - s \nabla E(W_n)$$