Session 9: Convex functions

Optimization and Computational Linear Algebra for Data Science

Léo Miolane

Contents

1. Recap of the videos
2. Convex sets and convex functions
3. Convex functions and derivatives
4. Jensen's inequality

Optimization

In machine learning, we often have to minimize functions

$$f(\theta) = \mathrm{Loss}(\mathrm{data}, \mathrm{model}_\theta) \quad \text{with respect to} \quad \underbrace{\theta \in \mathbb{R}^n}_{\text{model's parameters}}.$$

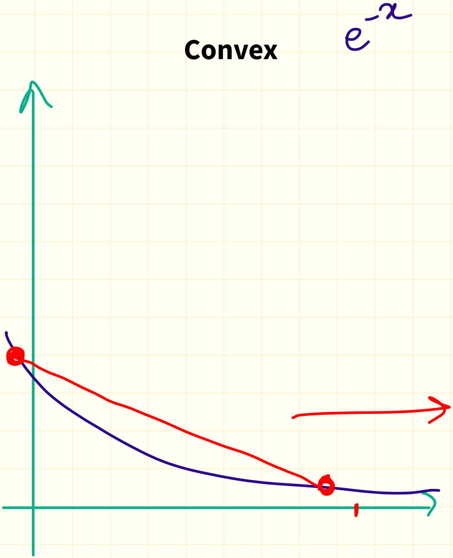
- For $n = 1, 2$, one could plot f to find the minimizer.
- This is intractable for larger dimension.

We will

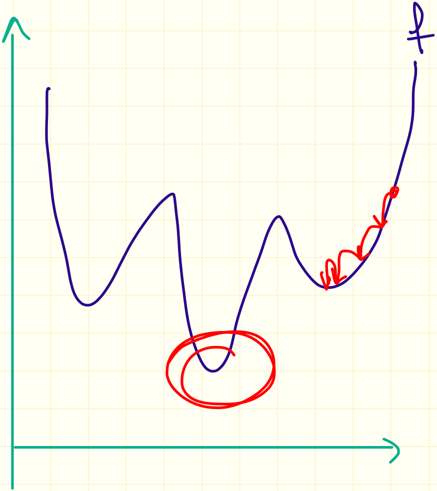
- focus on convex cost functions f.
- study gradient descent algorithms to minimize f.

Convex vs non-convex

Convex



Non-convex

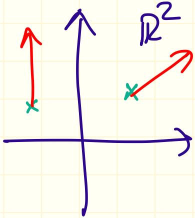


Gradient/Hessian

For $f : \mathbb{R}^n \to \mathbb{R}$:

- Gradient at $x \in \mathbb{R}^n$:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \in \mathbb{R}^n$$



- Hessian at $x \in \mathbb{R}^n$:

$$H_f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1\,\partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1}(x) & \frac{\partial^2 f}{\partial x_n\,\partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix} \in \mathbb{R}^{n\times n}$$

Taylor's formulas

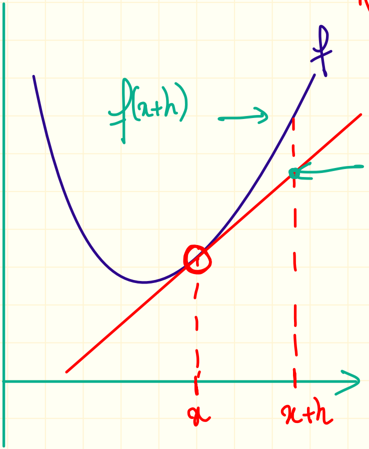
Let $x \in \mathbb{R}^n$. Heuristically, for $h \in \mathbb{R}^n$ "small", we have

$$f(x+h) \simeq f(x) + \langle \nabla f(x), h \rangle.$$

$u \cdot v$

$\langle u, v \rangle$ = dot product between u and v

not mathematically correct



$f(x) + \langle \nabla f(x), h \rangle$

"Rigorous Taylor"

$$\boxed{f(x+h) = f(x) + \nabla f(x) \cdot h + Err_1(h)}$$

$$\boxed{\frac{Err_1(h)}{\|h\|} \xrightarrow[h \to 0]{} 0}$$

0,001

Taylor's formulas

Let $x \in \mathbb{R}^n$. Heuristically, for $h \in \mathbb{R}^n$ "small", we have

$$f(x+h) \simeq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top H_f(x) h.$$

$$+ \ \mathrm{Err}_2(h)$$

$$\frac{\mathrm{Err}_2(h)}{\|h\|^2} \xrightarrow[h \to 0]{} 0$$

h^2

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x+h) \simeq f(x) + f'(x)h + \frac{1}{2} h f''(x) h$$

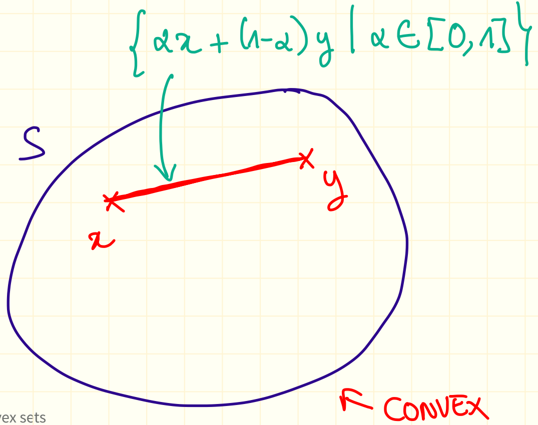
Convex sets

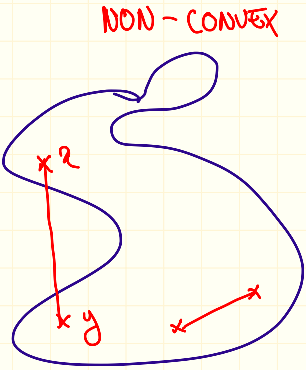
Convex set

Definition

A set $S \subset \mathbb{R}^n$ is called a convex set if for all $x, y \in S$ and all $\alpha \in [0, 1]$,

$$\alpha x + (1 - \alpha) y \in S.$$





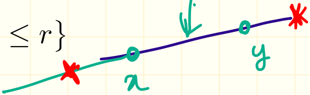
Exercise

1. Show that any subspace S of $\mathbb{R}^n$ is convex.
2. Let $\|\cdot\|$ be a (arbitrary) norm and $r \geq 0$. Show that the "ball" of radius r:

$$B(r) = \{x \in \mathbb{R}^n \mid \|x\| \leq r\}$$

is convex.

$\alpha x + (1-\alpha) y$



① Let $x, y \in S$ and $\alpha \in [0,1]$, since S is closed under linear combinations, $\alpha x + (1-\alpha) y \in S$, S is convex

② Let $x, y \in B(r)$, $\alpha \in [0,1]$

$$\|\alpha x + (1-\alpha) y\| \leq \|\alpha x\| + \|(1-\alpha) y\| = \alpha \|x\| + (1-\alpha)\|y\|$$

triangular ineq.

$$\leq \alpha r + (1-\alpha) r = r$$

$\longrightarrow \alpha x + (1-\alpha) y \in B(r)$.

Convex functions

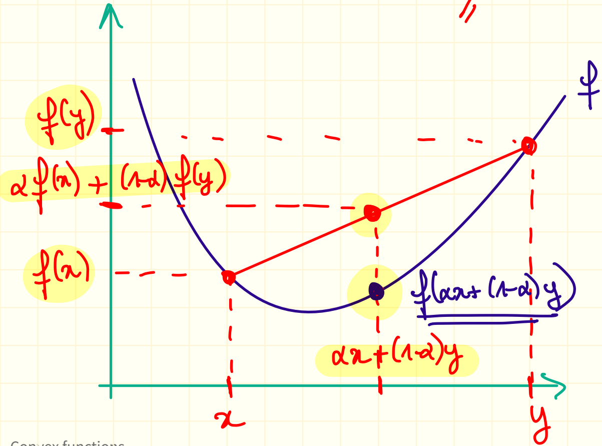
Convex / concave functions

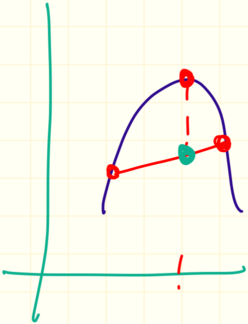
Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and all $\alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y). \tag{1}$$

$\geq$





Convex / concave functions

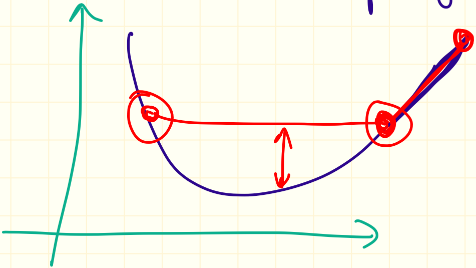
Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^n$ and all $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y). \tag{1}$$

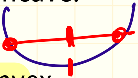
- We say that f is *strictly convex* is there is strict inequality in (1) whenever $x \neq y$ and $\alpha \in (0,1)$.
- A function f is called concave if $-f$ is convex.

$\longrightarrow$ inequality in the other direction.



Exercise

1. Show that any linear map $f : \mathbb{R}^n \to \mathbb{R}$ is convex and concave.
2. Show that a norm $\| \cdot \|$ is convex.
3. Show that the sum of two convex functions is also a convex function.



$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

① Let $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$

$$f(\alpha x + (1-\alpha) y) \overset{①}{\leq} \alpha f(x) + (1-\alpha) f(y)$$
$$\overset{}{=}$$
$$\overset{②}{\geq}$$

because f is linear

① → f is convex   ② → f is concave

② $\|\alpha x + (1-\alpha) y\| \leq \|\alpha x\| + \|(1-\alpha) y\| = \underbrace{\alpha}_{\geq 0} \|x\| + \underbrace{(1-\alpha)}_{\geq 0} \|y\|$

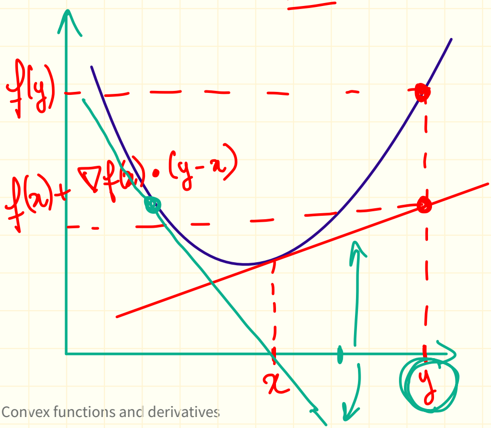
by trig. inequality.

Convex functions and derivatives

Proposition

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}^n$

$$f(\underbrace{y}_{x+h}) \geq f(x) + \langle \nabla f(x), \underbrace{(y - x)}_{h} \rangle.$$

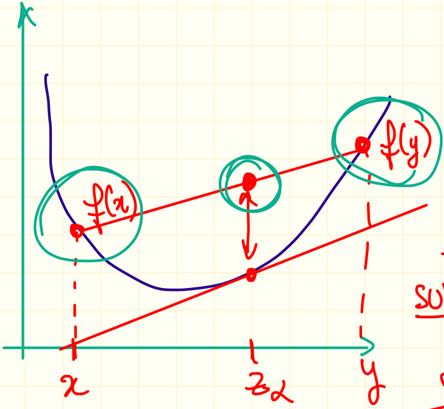


" The graph of a convex function is above its tangents "

Proof

$\Leftarrow$ Let's assume that for all $x, y \in \mathbb{R}^n$, $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$

Let $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$, define $z_\alpha = \alpha x + (1-\alpha) y$.



Applying the inequality twice:

$$\alpha f(x) \geq \alpha f(z_\alpha) + \nabla f(z_\alpha) \cdot (\alpha x - \alpha z_\alpha)$$

$$(1-\alpha) f(y) \geq (1-\alpha) f(z_\alpha) + \nabla f(z_\alpha) \cdot ((1-\alpha) y - (1-\alpha) z_\alpha)$$

SUM: $\alpha f(x) + (1-\alpha) f(y)$

$$\geq f(z_\alpha) + \cancel{\nabla f(z_\alpha) \cdot \underbrace{\left(\alpha x + (1-\alpha) y - z_\alpha\right)}_{=0}}$$

$\rightarrow$ f is convex

Proof

$\Rightarrow$) Let assume that f is convex.

Let $x, y \in \mathbb{R}^n$, let $t \in [0,1]$. write $\boxed{z_t = (1-t)x + ty}$.

$= x + t(y-x)$

Taylor's formula: $f(z_t) = f(x + \overbrace{t(y-x)}^{h})$

$= f(x) + t\nabla f(x)\cdot(y-x) + Err_1(t)$

f is convex $\quad f(z_t) \leqslant (1-t)f(x) + tf(y)$

Combining: $\quad tf(y) \geqslant tf(x) + t\nabla f(x)\cdot(y-x) + Err_1(t)$

$$\boxed{f(y) \geqslant f(x) + \nabla f(x)\cdot(y-x)} + \frac{Err_1(t)}{t}$$

$\xrightarrow{t \to 0} 0$

Proof

Minimizers of a convex function

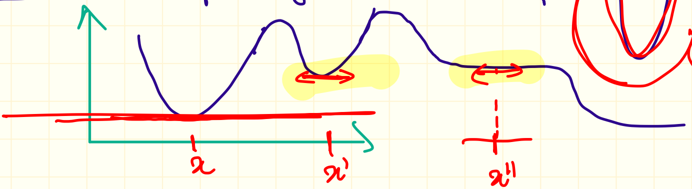
$$f(x+h) \simeq f(x) + \underbrace{\cancel{\nabla f(x)}}_{=0} \cdot h + \frac{1}{2} h^T Hf(x) h$$

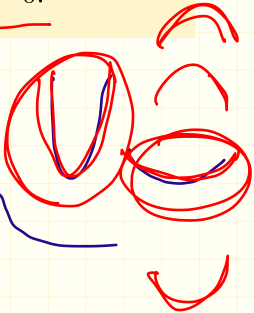
Corollary

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable $\boxed{\text{convex}}$ function and $x \in \mathbb{R}^n$.
Then

$$x \text{ is a minimizer of } f \iff \nabla f(x) = 0.$$

Proof: $\Rightarrow$) True for any differentiable f





$\Leftarrow$) Let $y \in \mathbb{R}^n$, $f(y) \geq f(x) + \underbrace{\nabla f(x)}_{=0} \cdot (y - x)$

$\geq f(x) \rightarrow x$ is a minimizer of f.

Hessian of convex function

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. Then f is convex if and only if for all $x \in \mathbb{R}^n$, $H_f(x)$ is positive semi-definite.

Recall :
- a matrix M is PSD if for all $v \in \mathbb{R}^n$, $v^T M v \geq 0$
- this is equivalent to saying that all the eigenvalues of M are ≥ 0

Example : $f(x) = \|x\|^2 = x_1^2 + \dots + x_n^2$

For all $x \in \mathbb{R}^n$, $H_f(x) = 2 I_{d_n} \leftarrow$ PSD hence f is convex

Hessian of convex function

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. Then f is convex if and only if for all $x \in \mathbb{R}^n$, $H_f(x)$ is positive semi-definite.

Remark: if $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable:

f is convex $\iff$ $f''(x) \geq 0$ for all $x \in \mathbb{R}$

Remark: If $H_f(x)$ is Positive definite for all $x \in \mathbb{R}^n$ then f is strictly convex. But the converse is not true in general.

Proof idea: $f(y) \underset{\geq}{\simeq} f(x) + \nabla f(x) \cdot (y-x) + \underbrace{\frac{(y-x)^T H_f(x) (y-x)}{2}}_{\geq 0}$

Jensen’s inequality

Jensen's inequality

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for all $x_1, \ldots, x_k \in \mathbb{R}^n$ and all $\alpha_1, \ldots, \alpha_k \geq 0$ such that $\sum_{i=1}^k \alpha_i = 1$ we have

concave

f("average") $\longrightarrow$

$$f\left(\sum_{i=1}^k \alpha_i x_i\right) \leq \sum_{i=1}^k \alpha_i f(x_i).$$

$\leftarrow$ "average of the f's"

$\geq$ $\leftarrow$ if f is concave

More generally, if X is a random variable that takes value in $\mathbb{R}^n$ we have

$$\boxed{f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].}$$

$\geq$

Example • $f: \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ is convex therefore,

for any random variable X, $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$

$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$

Example: entropy

- Let's consider a random variable X taking values in $\{1, \dots k\}$
- Let's write $p_i = \mathbb{P}(X = i)$ for $i \in \{1, \dots k\}$.

The entropy of X is defined as:

$$\boxed{H(X) = \sum_{i=1}^{k} p_i \log\Big(\underbrace{\frac{1}{p_i}}_{\geq 0}\Big)} \geq 0$$

log is a concave function (exercise!) hence, by Jensen:

$$H(X) = \sum_{i=1}^{k} \underbrace{p_i}_{\text{"}\alpha_i\text{"}} \overbrace{\log}^{g}\Big(\overbrace{\frac{1}{p_i}}^{\text{"}x_i\text{"}}\Big) \leq \overbrace{\log}^{g}\Big(\sum_{i=1}^{k} \underbrace{p_i}_{\text{"}\alpha_i\text{"}} \cdot \overbrace{\frac{1}{p_i}}^{\text{"}x_i\text{"}}\Big) = \log(k)$$

Example: entropy

$\longrightarrow 0 \leq H(X) \leq \log(k)$

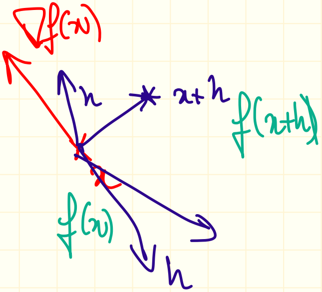
The value $\log(k)$ is achieved for $p_i = \frac{1}{k}$ for all i:

$$H(X) = \sum_{i=1}^{k} \frac{1}{k} \log\left(\frac{1}{\frac{1}{k}}\right) = \log(k)$$

The entropy is maximal when X is uniformly distributed over $\{1, \ldots k\}$

Questions?

$$f(x+h) \simeq f(x) + \langle h, \nabla f(x) \rangle$$



$$\max \{ \underline{v \cdot x} \mid \|v\| = 1 \}$$

Questions?

$$\#\text{paths}(k+1, i \to j) = \sum_{\ell \text{ neighbor of } j} \#\text{paths}(k, i \to \ell)$$

$$\frac{\nabla f(x)}{\|\nabla f(x)\|} = \underset{\substack{v \in \mathbb{R}^n \\ \|v\|=1}}{\operatorname{argmax}} \lim_{\substack{t \to 0 \\ >}} \boxed{\frac{f(x+tv) - f(x)}{t}}$$