CSED490Y: Optimization for Machine Learning

Week 03-2: Convex optimization

Namhoon Lee

POSTECH

Spring 2022

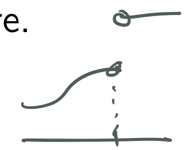
Convex function

(*) $\min_{x \in C} f(x)$: convex opt. f, C : convex

equiv.

Convex function?

- A C^0 function is convex if and only if the function is below its chord between any two points.
- A C^1 function is convex if and only if the function is above its tangent planes at any point.
- A C^2 function is convex if and only if it is curved upwards everywhere.



exp. ⇒ degree of differentiability.

C^0 func : i.e. continuous func.

C^1 func : " & differentiable

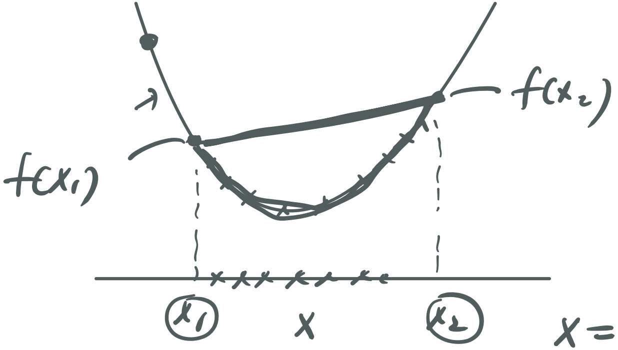
C^2 " : " & twice differentiable.

Convex function

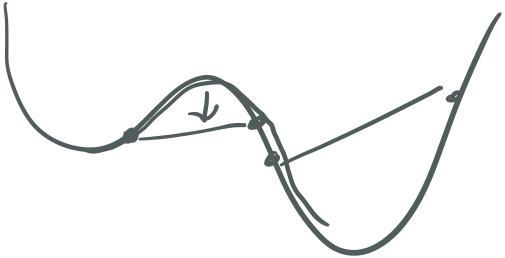
therefore.

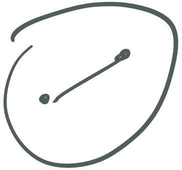
C^0 definition of "convex" functions : line/chords between any two points lies above (*)

convex



NOT





$x = \theta x_1 + (1-\theta) x_2 \qquad \theta \in [0,1]$

(*) $f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)$

$\forall \{x_1, x_2\} \in \mathbb{R}^d, \ \theta \in [0,1]$

Convex function

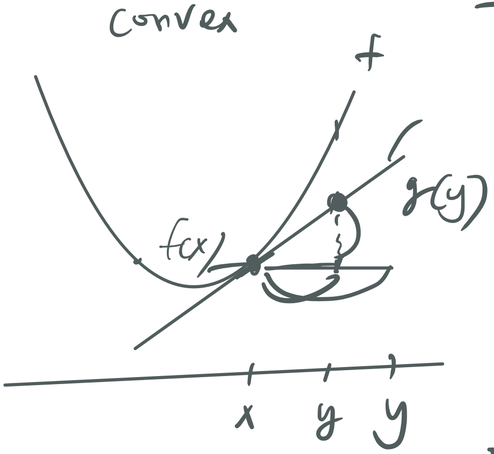
C^0 definition of convex functions

✓ $f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$, $\forall x_1, x_2 \in \mathbb{R}^n$ and $0 \leq \theta \leq 1$.

Convex function

f : differentiable ⊛ $x \in \mathrm{Dom} f$

C^1 definition of convex functions : func. always above the tangent plane at any point.



$$\boxed{\begin{aligned} f(y) &\geq g(y) \\ &= f(x) + \langle \nabla f(x), y-x \rangle \end{aligned}}$$

$g(x) = f(x)$

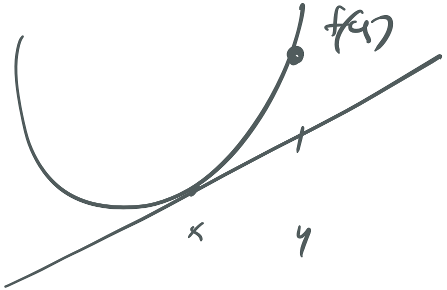
$g(y) - g(x) = \nabla g(x)^T (y-x)$

$\nabla g(x) = \nabla f(x)$

Convex function

C^1 definition of convex functions

$\underline{f}$ convex.



$$\circledast \qquad \underline{f(y)} \geq \underline{f(x)} + \langle \nabla f(x), y - x \rangle \ , \quad \forall \underline{x}, \underline{y} \in \mathbb{R}^n \ .$$

Convex function

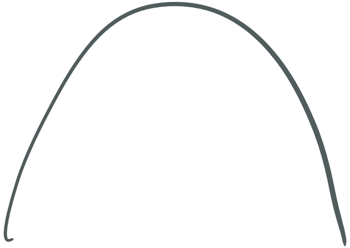
C^2 definition of convex functions : f curved upward

$x \in \mathbb{R}$

$\boxed{f''(x) \geq 0}$

f: convex

NOT



C^2 definition of convex functions

$$\nabla^2 f(x) = \begin{bmatrix} \ddots & \dfrac{\partial^2 f}{\partial x_i \partial x_j} & \\ & & \ddots \end{bmatrix}$$

$M \succeq 0$: M is positive semidefinite

symm.

all eigenvalues of M, λ_i

$i = \{1, \dots, n\}$

$\lambda_i \geq 0$

$\in \mathbb{R}^{n \times n}$

$$\underbrace{\nabla^2 f(x)}_{\boxed{\text{Hessian}}} \succeq 0\,, \quad \forall x \in \mathbb{R}^n\,.$$

Convex function

twice differentiable

Show that the $\boxed{C^2}$ definition is equivalent to the C^1 definition.

Convex function

Show that the C^2 definition is equivalent to the C^1 deifnition.

First, let's recall the fundamental theorem of calculus:

$\int_0^1 x^2 \, dx = \frac{1}{3} x^3 \Big|_0^1 = \frac{1}{3}$

$$\int_0^1 F'(t)dt = F(1) - F(0)$$

Convex function

Show that the C^2 definition is equivalent to the C^1 deifnition.

First, let's recall the fundamental theorem of calculus:

$$\int_0^1 F'(t)dt = F(1) - F(0)$$

Now consider the following:

$$\int_0^1 (x-y)^\top \nabla^2 f(\underbrace{tx + (1-t)y}_{})dt \; =$$

$\nabla f(\,\text{''}\,)$

Convex function

Show that the C^2 definition is equivalent to the C^1 deifnition.

First, let's recall the fundamental theorem of calculus:

$$\int_0^1 F'(t)dt = F(1) - F(0)$$

Now consider the following:

$(x-y)$

$$\int_0^1 (x-y)^\top \boxed{\nabla^2 f(tx + (1-t)y)} dt = \int_0^1 \frac{d}{dt}\Big(\nabla f(tx + (1-t)y)\Big) dt$$

Convex function

Show that the C^2 definition is equivalent to the C^1 deifnition.

First, let's recall the fundamental theorem of calculus:

$$\int_0^1 F'(t)dt = F(1) - F(0)$$

Now consider the following:

$$\int_0^1 (x-y)^\top \nabla^2 f(tx + (1-t)y)dt = \int_0^1 \frac{d}{dt}\Big(\nabla f(tx + (1-t)y)\Big)dt = \nabla f(x) - \nabla f(y)$$

Convex function

Multiplying by $x - y$ both sides gives

$$\int_0^1 (x-y)^\top \nabla^2 f(tx + (1-t)y)(x-y)dt = \langle \nabla f(x) - \nabla f(y), x - y \rangle$$

Convex function

$v^\top M v \geq 0 \quad \forall v \in \mathbb{R}^d$

Multiplying by $x - y$ both sides gives

$$\underbrace{\int_0^1 (x-y)^\top \nabla^2 f(tx + (1-t)y)(x-y)dt}_{\geq 0} = \langle \nabla f(x) - \nabla f(y), x - y\rangle$$

By applying C^2 definition, we obtain

$$(*) \qquad \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$$

Multiplying by $x - y$ both sides gives

$$\int_0^1 (x-y)^\top \nabla^2 f(tx+(1-t)y)(x-y)dt = \langle \nabla f(x) - \nabla f(y), x-y \rangle$$

$\boxed{\forall \{x, y\}}$

$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle$

$\cancel{f(y)+f(x)} \geq \cancel{f(x)+f(y)} + \langle \nabla f(x), y-x \rangle + \langle \nabla f(y), x-y \rangle$

By applying C^2 definition, we obtain

monotone : $\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq 0$

- It's called function is monotone; *i.e.*, C^2 function is monotone.
- You can also show that C^1 function is monotone.

Convex function

Next, consider the following:

$$\int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt$$

$$\frac{d}{dt} f((y-x)t+x)$$

Convex function

$\int_0^1 F'(t)\,dt = F(1) - F(0)$

Next, consider the following:

$$\int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt = \int_0^1 \frac{d}{dt}\Big(f((y-x)t+x)\Big)dt \quad =$$

Convex function

Next, consider the following:

$$\int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt = \int_0^1 \frac{d}{dt}\Big(f((y-x)t+x)\Big)dt = f(y)-f(x)$$

Convex function

Next, consider the following:

$$\int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt = \int_0^1 \frac{d}{dt}\Big(f((y-x)t+x)\Big)dt = f(y) - f(x)$$

Rearranging it gives

h(t) h(0) = ∇f(x)

$$f(y) = f(x) + \int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt$$

C' def. $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

Convex function

C_1 def.

Next, consider the following:

$$\int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt = \int_0^1 \frac{d}{dt}\Big(f((y-x)t+x)\Big)dt = f(y)-f(x)$$

Rearranging it gives

$\int_0^1 h(t)\,dt \geq h(\cdot)$

$$(*) \quad f(y) = f(x) + \int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt$$

$= \langle \nabla f(x), y-x \rangle$

We want to relate this to the C^1 definition, while using monotonicity.

def.

$(*)$

$\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq 0$

Convex function

$\in [0,1]$

From the monotonicity, we can show that the integrand is smallest at $t = 0$, *i.e.*,

$$(*) \quad \langle \nabla f((y-x)t + x) - \nabla f(x), (y-x)t + \cancel{x} - \cancel{x} \rangle \geq 0$$

$$(*) \quad \langle \nabla f((y-x)t + x) - \nabla f(x), y - x \rangle \geq 0$$

$$= h(t) - h(0) \geq 0 \iff h(t) \geq h(0)$$

$$\int_0^1 h(t) \geq h(0) \cdot 1$$

Convex function

From the monotonicity, we can show that the integrand is smallest at $t = 0$, *i.e.*,

$$\begin{aligned}\langle \nabla f((y-x)t+x) - \nabla f(x), (y-x)t+x-x \rangle &\geq 0 \\ \langle \nabla f((y-x)t+x) - \nabla f(x), y-x \rangle &\geq 0\end{aligned}$$

Therefore, we can say

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$$

Convex function

From the monotonicity, we can show that the integrand is smallest at $t = 0$, *i.e.*,

$$\begin{aligned} \langle \nabla f((y-x)t+x) - \nabla f(x), (y-x)t+x-x \rangle &\geq 0 \\ \langle \nabla f((y-x)t+x) - \nabla f(x), y-x \rangle &\geq 0 \end{aligned}$$

Therefore, we can say

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$$

- If function is monotone, it's convex.
- More rigorous proofs exist.

$v^T M v \geq 0$

Example: For $f(x) = x^\top Qx$ where Q is postive semidefinite, show f is convex using definitions of convex functions.

① C^2 def: $\nabla f(x) = 2Qx$

$\nabla^2 f(x) = 2Q \succeq 0$ yes.

$\therefore f$ convex.

② C^1 def: $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$

$y^T Q y \overset{?}{\geq} x^T Q x + \langle 2Qx, y-x \rangle$

$y^T Q y \geq x^T Q x + 2x^T Q y - 2x^T Q x$

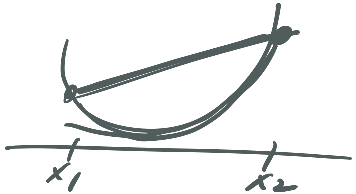
$y^T Q y - 2x^T Q y + x^T Q x \overset{?}{\geq} 0$

$(y-x)^T Q (y-x) \overset{?}{\geq} 0$

p.s.d.

③ ϵ def,

C^0 def.:



$$\boxed{f(\theta x_1 + (1-\theta)x_2) \le \theta f(x_1) + (1-\theta) f(x_2)}$$

Example: Show p-norm is convex.

$f(x) = \|x\|_p$

? ✓

$$\|\theta x_1 + (1-\theta)x_2\|_p \le \theta \|x_1\|_p + (1-\theta)\|x_2\|_p$$

trick

triangle ineq.

$\theta \in [0,1]$

$$\|\theta x_1 + (1-\theta)x_2\|_p \le \|\theta x_1\|_p + \|(1-\theta)x_2\|_p$$
$$= \theta \|x_1\|_p + (1-\theta)\|x_2\|_p$$

same

Convex function

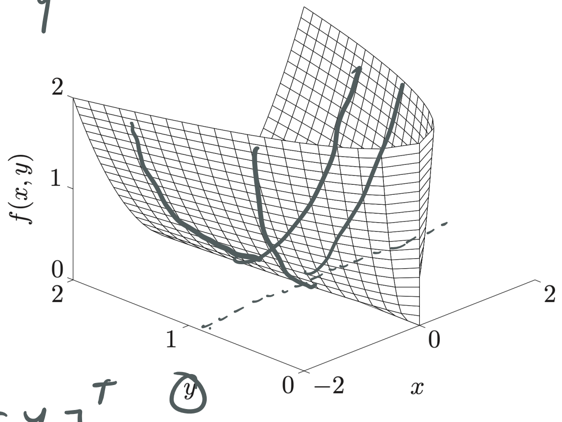
$f(x, y=1) = x^2$

Example: Show $f(x,y) = x^2/y$ is convex.

$\frac{2x}{y}$ $\qquad$ $\frac{x^2}{y^2}$

C^2 def.

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x \partial x} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y \partial y} \end{bmatrix} = \begin{bmatrix} \frac{2}{y} & -\frac{2x}{y^2} \\ \frac{-2x}{y^2} & \frac{2x^2}{y^3} \end{bmatrix}$$



$$= \frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix} = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \succeq 0 \qquad \forall y \geq 0$$

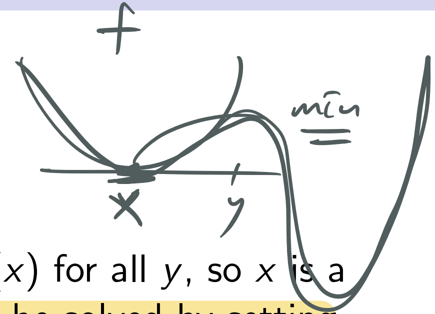
Convex function

Another way to show a function is convex is through convexity-preserving operations:

- Nonnegative weighted sum; *i.e.* if $\alpha, \beta \geq 0$ and f_1, f_2 convex, $\alpha f_1 + \beta f_2$ is convex.
- Pointwise maximum; *i.e.*, if $f_1, \ldots, f_m$ are convex, $\max\{f_1(x), \ldots, f_m(x)\}$ is convex.
- Composition with affine map; *i.e.*, if f is convex, $f(Ax + b)$ is convex.

Convex function

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$



More on convex functions..

- Notice from C^1 definition that $\nabla f(x) = 0$ implies $f(y) \geq f(x)$ for all y, so x is a global minimizer; this further explains why least squares can be solved by setting the derivative equal to zero.
- Strictly-convex function have at most one global minimum; w and v can't both be global minima if $w \neq v$; it would imply convex combinations u of w and v would have $f(u)$ below the global minimum.

Convex function

For strictly convex objective f there can be at most one global optimum.

Convex function

For strictly convex objective f there can be at most one global optimum.

Proof:

1. Suppose x^* is a local minimum and also there exists another local minimum $x^{\#}$ $(\neq x^*)$.

Convex function

For strictly convex objective f there can be at most one global optimum.

Proof:

1. Suppose x^* is a local minimum and also there exists another local minimum $x^\#$ $(\neq x^*)$.
2. Since f is convex (because it is strictly convex), $f(x^*)$ and $f(x^\#)$ are both global minima, and $f(x^*) = f(x^\#)$.

Convex function

For strictly convex objective f there can be at most one global optimum.

Proof:

1. Suppose x^* is a local minimum and also there exists another local minimum $x^\#$ $(\neq x^*)$.
2. Since f is convex (because it is strictly convex), $f(x^*)$ and $f(x^\#)$ are both global minima, and $f(x^*) = f(x^\#)$.
3. The C^0 definition for $y = \theta x^* + (1-\theta)x^\#$, *i.e.*,

$$f(y) < \theta f(x^*) + (1-\theta)f(x^\#) = \theta f(x^*) + (1-\theta)f(x^*) = f(x^*)$$

   contradicts that x^* is a global minimum.

Convex function

For strictly convex objective f there can be at most one global optimum.

Proof:

1. Suppose x^* is a local minimum and also there exists another local minimum $x^\#$ $(\neq x^*)$.
2. Since f is convex (because it is strictly convex), $f(x^*)$ and $f(x^\#)$ are both global minima, and $f(x^*) = f(x^\#)$.
3. The C^0 definition for $y = \theta x^* + (1-\theta)x^\#$, *i.e.*,

   $$f(y) < \theta f(x^*) + (1-\theta)f(x^\#) = \theta f(x^*) + (1-\theta)f(x^*) = f(x^*)$$

   contradicts that x^* is a global minimum.
4. This means that for $x^\#$ to be a local minimum, it must be that $x^\# = x^*$.

Thank you

Any questions?

Credits

A lot of material in this course is borrowed or derived from the following:

- Numerical Optimization, Jorge Nocedal and Stephen J. Wright.
- Convex Optimization, Stephen Boyd and Lieven Vandenberghe.
- Convex Optimization, Ryan Tibshirani.
- Optimization for Machine Learning, Martin Jaggi and Nicolas Flammarion.
- Optimization Algorithms, Constantine Caramanis.
- Advanced Machine Learning, Mark Schmidt.