Optimization for Data Science

F. Rinaldi[1]

1 Università degli Studi di Padova | Dipartimento Matematica

Padova
2020

Outline

Optimization for Data Science

Convex Sets

Definition

A set C is said *convex* if, for any two points $x, y \in C$ the segment connecting the two points is contained in C. That is, if for every point

$$z = \lambda x + (1 - \lambda)y,$$

with $x, y \in C$ and $\lambda \in [0, 1]$, we have $z \in C$.

Proposition (Intersection)

The intersection of convex sets is convex.

Convex Sets II

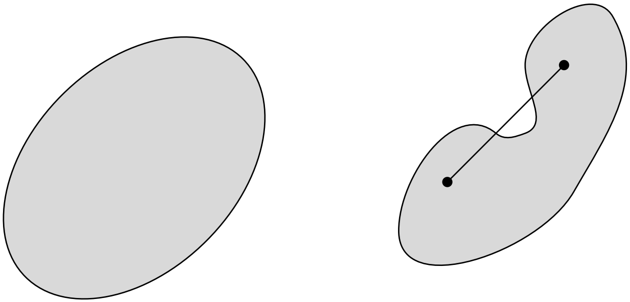

Figure: Example of a convex (left) and non-convex set (right).

Some useful definitions

Convex Combination

A point $x \in \mathbb{R}^n$ is a *convex combination* of points $v_1, \ldots, v_p \in \mathbb{R}^n$ if we have:

$$\begin{aligned} x &= \sum_{i=1}^{p} \alpha_i \cdot v_i, \\ & \sum_{i=1}^{p} \alpha_i = 1, \\ & \alpha_i \geq 0,\ i = 1, \ldots,\ p. \end{aligned}$$

Proper Convex Combination

A point $x \in \mathbb{R}^n$ is a *proper convex combination* of points $v_1, \ldots, v_p \in \mathbb{R}^n$ if we further have $\alpha_i \in (0, 1)$ for all $i \in \{1, \ldots, p\}$.

Some useful definitions II

Convex Hull

The set of all convex combinations of points $v_1, \ldots, v_p \in \mathbb{R}^n$ is usually called *convex hull* of $v_1, \ldots, v_p$ and usually denoted by $conv(v_1, \ldots, v_p)$.

Extreme Point

Given a convex set $C \subseteq \mathbb{R}^n$ a point $\bar{x} \in C$ is an *extreme point* of C if $\bar{x}$ cannot be given as a proper convex combination of two points in C. That is, if we cannot find two points $y, z \in C$ such that $\bar{x} = \alpha y + (1 - \alpha) z$, with $\alpha \in (0, 1)$.

Convex functions

Convex Function

Let $C \subseteq \mathbb{R}^n$ be a convex set. Let $f : C \to \mathbb{R}$ be a function defined in C. We say that f is *convex* over C if for any $y, z \in C$ and $\lambda \in [0, 1]$ we have

$$f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z);$$

we say that f is *strictly convex* over C if for any $y, z \in C$, with $y \neq z$, and $\lambda \in (0, 1)$ we have

$$f(\lambda y + (1-\lambda)z) < \lambda f(y) + (1-\lambda)f(z).$$

Convex functions

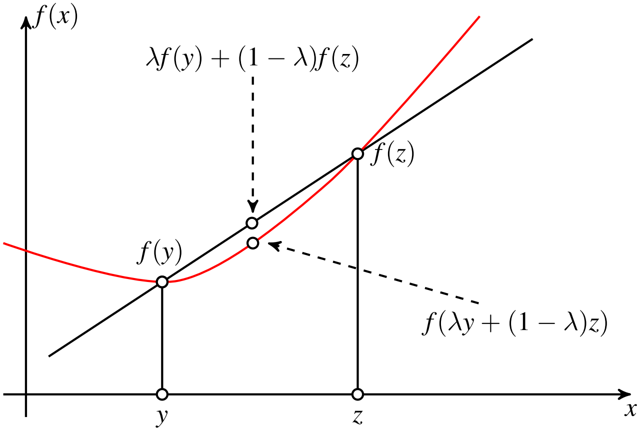

Figure: Example of convex function (chosen points are y and z)

Properties of Convex Functions

Proposition (Convexity of the Epigraph)

Let $C \subseteq \mathbb{R}^n$ be a convex set and $f : C \to \mathbb{R}$ be a function defined in C. We say that f is *convex* over C if and only if the epigraph of f

$$\text{epi}(f) = \{(x, \alpha) \in C \times \mathbb{R} : \ \alpha \geq f(x)\}$$

is a convex set.

Operations Preserving Convexity I

Proposition (Non-negative Weighted Sum)

Let $C \subseteq \mathbb{R}^n$, be a convex set and $f_i : C \to \mathbb{R},\ i = 1, \ldots, m$ be convex functions defined in C. We have that the function

$$f(x) = \sum_{i=1}^{m} \alpha_i f_i(x),$$

with $\alpha_i \geq 0,\ i = 1, \ldots, m$ is a *convex* function over C. Furthermore, if there also exists an index i s.t. $\alpha_i > 0$ and f_i strictly convex over C, then f is strictly convex over C.

Operations Preserving Convexity II

Proposition (Pointwise Maximum)

Let $C \subseteq \mathbb{R}^n$, be a convex set and $f_i : C \to \mathbb{R},\ i = 1, \ldots, m$ be convex functions defined in C. We have that the function

$$f(x) = \max_{i \in I} f_i(x),$$

with $I = \{1, \ldots, m\}$ a set of indices, is a convex function over C. Furthermore, if f_i, with $i = 1, \ldots, m$ are strictly convex over C, then f is strictly convex over C.

Operations Preserving Convexity III

Proposition (Composition)

Let $C \subseteq \mathbb{R}^n$, be a convex set, $g : C \to \mathbb{R}$ be a convex function defined in C and $h : conv(g(C)) \to \mathbb{R}$ be a convex and non-decreasing function over the convex hull of points $g(C) = \{\alpha \in \mathbb{R} : \alpha = g(x),\ x \in C\}$. Then the function

$$f(x) = h(g(x))$$

is convex over C. Furthermore, if g is strictly convex over C and h is strictly convex and non-decreasing over $conv(g(C))$, f is strictly convex over C.

Operations Preserving Convexity IV

Proposition (Linear Transformation)

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a convex function defined in $\mathbb{R}^m$. Let A be an $m \times n$ real matrix. We have that the function

$$F(x) = f(Ax),$$

is convex over $\mathbb{R}^n$.

Gradient and Hessian

Definition (Gradient and Hessian)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a given function. We call *gradient* of f at x, the following vector:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \dots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix},$$

where $\frac{\partial f(x)}{\partial x_i}$ indicates the partial derivative with respect to the i-th component of x. We call *Hessian* of f at x, the following matrix:

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \dots & \dots & \dots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_n x_n} \end{pmatrix},$$

where $\frac{\partial^2 f(x)}{\partial x_i x_j}$ indicates the second partial derivative with respect to the i-th and j-th component of x.

Closed and Open Sets

Definition (Closed and Open Sets)

First we say that x is a *closure point* of a subset $C \in \mathbb{R}^n$ if there exists a sequence of points in C converging to x. The *closure* is denoted by $\mathrm{cl}(C)$. A set is *closed* if it is equal to its closure. A set is *open* if its complement $\{x \notin C\}$ is closed.

Definition (Directional Derivative)

Directional derivative of f along d has the following form

$$\lim_{h \to 0^+} \frac{f(x+hd)-f(x)}{h} = \nabla f(x)^\top d.$$

First Order Convexity

Proposition (First Order Convexity Condition)

Let $C \subseteq \mathbb{R}^n$ be an open convex set. If f is continuously differentiable over C then:

(i) f is convex over C if and only if for all $y, z \in C$ we have:

$$f(z) - f(y) \geq \nabla f(y)^\top (z - y);$$

(ii) f is strictly convex over C if and only if for any $y, z \in C$, with $y \neq z$, we have:

$$f(z) - f(y) > \nabla f(y)^\top (z - y).$$

First Order Convexity Conditions

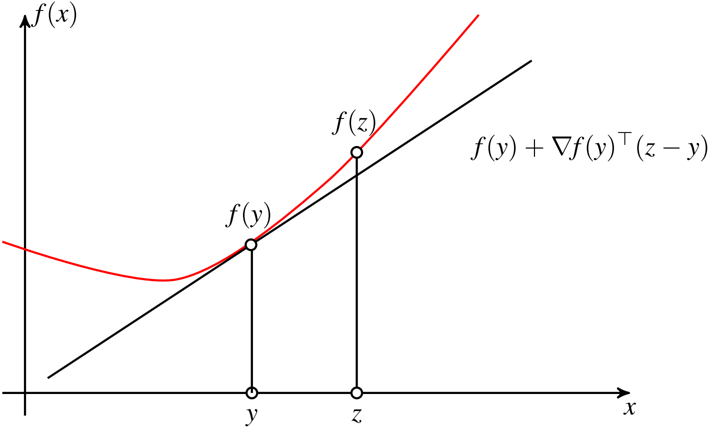

Figure: Convexity conditions (i) and (ii) (chosen points are y and z)

Second Order Convexity

Proposition (Second Order Convexity Condition)

Let $C \subseteq \mathbb{R}^n$ be an open convex set. If f is twice continuously differentiable over C then:

(iii) f is convex over C if and only if for all $x \in C$ we have:

$$d^\top \nabla^2 f(x) d \geq 0 \quad \forall\, d \in \mathbb{R}^n.$$

(iv) f is strictly convex over C if for all $x \in C$ we have:

$$d^\top \nabla^2 f(x) d > 0 \quad \forall\, d \in \mathbb{R}^n, d \neq 0.$$

Comments on the Results

As we will see later on,

- Properties (i) and (ii) are relevant when studying the minimum points of this particular class of functions.
- Properties (iii) and (iv) are also useful since enable us to understand when a given function is convex.

Remark

Notice anyway that positive definiteness of the Hessian is not a necessary condition for ensuring strict convexity.

Example

A simple example is the function $y = x^4$ with $x = 0$. Indeed, $f(x) = x^4$ is strictly convex but $\frac{\partial^2 f(0)}{\partial x^2} = 0$.

Comments on the Results II

Proposition

Let Q be an $n \times n$ real symmetric matrix. If f is defined as

$$f(x) = x^\top Qx + c^\top x.$$

Then, we have

- f is convex if and only if Q is positive semidefinite;
- f is strictly convex if and only if Q is positive definite.

Convex programming problem

Convex programs

A *convex programming problem* is a problem of the following form:

$$\begin{array}{ll} \min & f(x) \\ s.t. & x \in C \end{array}$$

where C is a convex set and f is convex over C.

Global and Local Minima

Global Minimum

A point $x^* \in C$ is a *global minimum* of f over C, if

$$f(x^*) \leq f(x), \quad \text{for all } \; x \in C.$$

Local Minimum

A point $x^* \in C$ is a *local minimum* of f over C, if there exists a neighborhood $B(x^*;\rho) = \{x \in \mathbb{R}^n : \; \|x - x^*\| < \rho\}$, with $\rho > 0$ s.t.

$$f(x^*) \leq f(x), \quad \text{for all } \; x \in C \cap B(x^*;\rho).$$

Global and Local Minima II

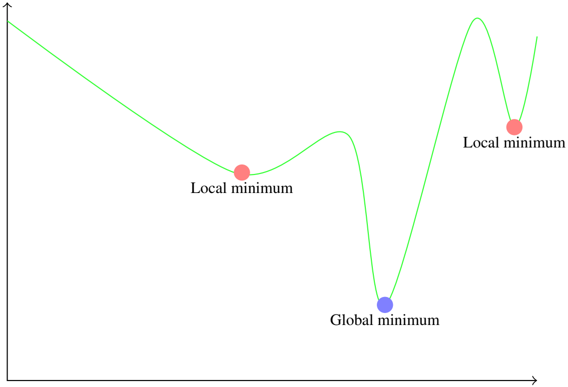

Figure: Examples of global and local minima

Comments on Global and Local Minima

- Finding a global minimum much harder than finding a local one!!!
- Algorithms that find local minima are cheaper than algorithms that find global minima.
- When dealing with huge scale data better stay local!

Equivalence between global and local minima

Proposition [Equivalence between global and local minima]

Let $C \subseteq \mathbb{R}^n$ be a convex set and f a convex function over C. Then every local minimum of f over C is also a global minimum. Furthermore, the set $X^* \subseteq C$ of global minima for f over C is a convex set.

Proof of the Equivalence

Let x^* be a local minimum of f over C. From definition of local minimum, we can define an open ball $B(x^*;\rho)$ with $\rho > 0$ such that

$$f(x^*) \leq f(y), \;\; \forall\, y \in B(x^*;\rho) \cap C. \tag{1}$$

Let x be any point in C. From convexity of C we have that

$$z(\lambda) = (1-\lambda)x^* + \lambda x \in C, \qquad \text{for all} \qquad \lambda \in [0,1].$$

Since, when $\lambda = 0$, we have $z(0) = x^*$, we can find a value $\bar{\lambda} \in (0,1]$ such that

$$z(\bar{\lambda}) = (1-\bar{\lambda})x^* + \bar{\lambda}x \in B(x^*;\rho) \cap C.$$

Then using (1), we get:

$$f(x^*) \leq f(z(\bar{\lambda})).$$

Proof of the Equivalence II

Using convexity of objective function, we can write:

$$f(x^*) \leq f(z(\bar{\lambda})) = f((1-\bar{\lambda})x^* + \bar{\lambda}x) \leq (1-\bar{\lambda})f(x^*) + \bar{\lambda}f(x),$$

which, by taking into account that $\bar{\lambda} > 0$, implies:

$$f(x^*) \leq f(x). \tag{2}$$

Since x is arbitrarily chosen in C, from (2) we have that x^* is a global minimum.

Proof of the Equivalence III

Now we prove the second part of the theorem. In case, $X^* = \emptyset$ or $X^* = \{x^*\}$ we immediately get the result. Let us consider two different points $x^*, y^* \in X^*$, i.e.

$$f(x^*) = f(y^*) = \min_{x \in C} f(x).$$

Taking into account that C is a convex set and f is convex over C, for all $\lambda \in [0, 1]$, we have

$$f((1-\lambda)x^* + \lambda y^*) \leq (1-\lambda)f(x^*) + \lambda f(y^*) = f(x^*),$$

which implies $conv(x^*, y^*) \subseteq X^*$ and X^* is convex.

Uniqueness of global minimum

In the next proposition we state another interesting equivalence for f strictly convex function over C.

Proposition [Uniqueness of global minimum]

Let $C \subseteq \mathbb{R}^n$ be a convex set and f a strictly convex function over C. Then if x^* is local minimum of f over C it is also the only local and global minimum for f over C.

Summarizing

- Convex sets and convex functions
- Basic operations preserving convexity
- First/Second order convexity conditions
- Convex programming problems
- Equivalence between local and global minima in convex problems

Convex is Cool

Strong Convexity

- Strong convexity is an important concept in optimization.
- Strong convexity is used for proving linear convergence rate of many gradient descent based algorithms.
- In this notes we present some useful results on strong convexity.

Definition

Let $f : C \to \mathbb{R}$ be a continuous function and σ a positive scalar. We say that f is *strongly convex* over C with coefficient σ if for all $x, y \in C$ and all $\lambda \in [0, 1]$ we have:

$$f(\lambda x + (1 - \lambda)y) + \frac{\sigma}{2}\lambda(1 - \lambda)\|x - y\|^2 \leq \lambda f(x) + (1 - \lambda)f(y).$$

Strong Convexity

- It can be seen as a kind of "parameterized strict convexity"
- We call $z(\lambda) = \lambda x + (1-\lambda)y$
- By taking into account the definition, we can write for all $x, y \in C$ and all $\lambda \in [0, 1]$:

$$[\lambda f(x) + (1-\lambda)f(y)] - f(z(\lambda)) \geq \frac{\sigma}{2}\lambda(1-\lambda)\|x-y\|^2.$$

Comparison with strictly convex functions

- lhs is > 0 when dealing with strictly convex functions
- lhs is lower bounded by a term that depends on $\|x-y\|$ and σ!!!

Strong Convexity II

Then f is strictly convex over C. We have that there exists a unique minimizer $x^* \in C$ and by setting $y = x^*$ in the definition, we can write

$$f(\lambda x + (1-\lambda)x^*) + \frac{\sigma}{2}\lambda(1-\lambda)\|x - x^*\|^2 \leq \lambda f(x) + (1-\lambda)f(x^*).$$

Now, if we properly rewrite the right-hand side, we get

$$f(\lambda x + (1-\lambda)x^*) + \frac{\sigma}{2}\lambda(1-\lambda)\|x - x^*\|^2 \leq f(x^*) + \lambda(f(x) - f(x^*)).$$

By properly rewriting the expression above and dividing by $\lambda \in (0,1)$, we have

$$\frac{\sigma}{2}(1-\lambda)\|x-x^*\|^2 \leq \frac{f(\lambda x + (1-\lambda)x^*) - f(x^*)}{\lambda} + \frac{\sigma}{2}(1-\lambda)\|x-x^*\|^2 \leq f(x) - f(x^*).$$

Finally, by taking limit $\lambda \to 0$, we can write

$$f(x) \geq f(x^*) + \frac{\sigma}{2}\|x - x^*\|^2, \quad \forall\, x \in C.$$

Properties of Strong Convexity Functions

Proposition (Equivalent Notions of Strong Convexity)

Let f be a continuously differentiable function. Then, the following conditions are equivalent:

(*i*) f is strongly convex with coefficient σ over C;

(*ii*) $g(x) = f(x) - \frac{\sigma}{2}\|x\|^2$ is a convex function over C;

(*iii*) $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \sigma \|x - y\|^2, \quad \forall x, y \in C$;

(*iv*) $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\sigma}{2}\|x - y\|^2, \quad \forall x, y \in C.$

If f is twice continuously differentiable, then we have that the above properties are equivalent to:

(*v*) matrix $\nabla^2 f(x) - \sigma I$ is positive semidefinite for every $x \in \text{int}(C)$, where I is the identity matrix.

Results implied by Strong Convexity

Proposition (Other Results)

Let f be a continuously differentiable function. If f is strongly convex with coefficient σ over C, then we have:

(i) $\frac{1}{2}\|\nabla f(x)\|^2 \geq \sigma(f(x) - f(x^*))$ with x^* minimum for f over C;

(ii) $\|\nabla f(x) - \nabla f(y)\| \geq \sigma\|x - y\|, \quad \forall x, y \in C$;

(iii) $f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\sigma}\|\nabla f(x) - \nabla f(y)\|^2, \quad \forall x, y \in C$;

(iv) $(\nabla f(x) - \nabla f(y))^\top (x - y) \leq \frac{1}{\sigma}\|\nabla f(x) - \nabla f(y)\|^2, \quad \forall x, y \in C$.

(i) is usually called *Polyak-Lojasiewicz inequality*.