Lecture 1

5th of October 2018

A. Frangioni

1.1 Unconstrained optimization

Until now we stated that the best conditions for finding the optimum are encountered when the domain is a compact set and we have many derivatives.

Now we need to consider when we can stop our algorithm.

Definition 1.1.1 (Unconstrained optimization problem)**.** *We term **unconstrained optimization problem** the following*

$$(P)\ f_* = \min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \mathbb{R}^n\}$$

In unconstrained optimization, we deal with an unbounded set ($\mathbb{R}^n$), hence Weierstrass theorem does not apply. Because of this reason, we have no guarantee that a minimum $\mathbf{x}_*$ exists; moreover, provided its existence, finding it is an NP-hard problem. In order to make things easier, in practice we use a weaker condition: **local minimality**.

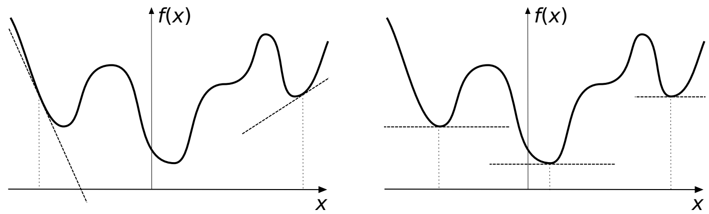
Definition 1.1.2 (Local minimum)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. $\mathbf{x}_*$ is a **local minimum** if it is a global minimum in a ball around $\mathbf{x}^*$. Formally,*

$$\min\{f(\mathbf{x}) \ : \ \mathbf{x} \in \mathcal{B}(\mathbf{x}_*, \varepsilon)\}$$

for some $\varepsilon > 0$.

*Also, $\mathbf{x}^*$ is a **strict local minimum** if $f(\mathbf{x}) < f(\mathbf{x}')$ $\forall \mathbf{x}' \in \mathcal{B}(\mathbf{x}_*,\ \varepsilon)$.*

If $f'(\mathbf{x}) < 0$ or $f'(\mathbf{x}) > 0$, $\mathbf{x}$ clearly cannot be a local minimum, as shown in Figure 1.1.



(a) Points where the derivative is not 0.

(b) Points where the derivative is 0.

FIGURE 1.1: Minima and not minima.

1.1.1 First order model

Do you recall?

The first order model of f is $L_{\mathbf{x}}(\mathbf{x}') = f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{x}' - \mathbf{x})$, such that $\forall \mathbf{x}' \in \mathbb{R}^n$ close to $\mathbf{x}$ $f(\mathbf{x}') = L_{\mathbf{x}}(\mathbf{x}') + R(\mathbf{x}' - \mathbf{x})$, where $R(\cdot)$ is called **residual** and it has the property of quadratic convergence: $\lim\limits_{\|h\| \to 0} \frac{R(h)}{\|h\|} = 0$.

Fact 1.1.1. *Let f be differentiable, if $\mathbf{x}$ is a local minimum, then $\nabla f(\mathbf{x}) = 0$.*

In optimization, we are interested in moving towards a (local) minimum $\mathbf{x}_*$ as fast as possible.

Fact 1.1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be the objective function of the optimization problem (P) and let $\mathbf{x} \in \mathbb{R}^n$ be the current point at a generic iteration. In order to get closer to the optimum, we need to take a step along the* anti-gradient *direction. Formally, $\mathbf{x}(\alpha) = \mathbf{x} - \alpha \nabla f(\mathbf{x})$, where $\alpha \in \mathbb{R}$ is called* step size.

Proof by contraddiction. Let us assume that $\mathbf{x}$ is a local minimum but $\nabla f(\mathbf{x}) \neq 0$.

In our case, $\mathbf{x}' = \mathbf{x} - \alpha \nabla f(\mathbf{x})$, so let us plug it into the *remainder version* of the Tailor's first-order model:

$$\begin{aligned} f(\mathbf{x}') &=< \nabla f(\mathbf{x}), \mathbf{x}' - \mathbf{x} > + f(\mathbf{x}) + R(\mathbf{x}' - \mathbf{x}) \\ &=< \nabla f(\mathbf{x}), \mathbf{x} - \alpha \nabla f(\mathbf{x}) - \mathbf{x} > + f(\mathbf{x}) + R(\mathbf{x} - \alpha \nabla f(\mathbf{x}) - \mathbf{x}) \\ &=< \nabla f(\mathbf{x}), -\alpha \nabla f(\mathbf{x}) > + f(\mathbf{x}) + R(-\alpha \nabla f(\mathbf{x})) \\ &= f(\mathbf{x}) - \alpha \left\| \nabla f(\mathbf{x}) \right\|^2 + R(-\alpha \nabla f(\mathbf{x})) \end{aligned}$$

Once we fixed the moving direction, we can choose the step size α, so it can be proved that $\lim\limits_{\alpha \to 0} \frac{R(-\alpha \nabla f(\mathbf{x}))}{\|\alpha \nabla f(\mathbf{x})\|} = 0$, that is equivalent by definition to $\forall \varepsilon > 0 \, \exists \bar{\alpha} > 0$ s.t. $\frac{R(-\alpha \nabla f(\mathbf{x}))}{\alpha \|\nabla f(\mathbf{x})\|} \leq \varepsilon \; \forall \; 0 \leq \alpha < \bar{\alpha}$.

If we take $\varepsilon < \|\nabla f(\mathbf{x})\|$, we get $R(-\alpha\nabla f(\mathbf{x})) < \alpha \|\nabla f(\mathbf{x})\|^2$, then

$$f(\mathbf{x}(\alpha)) = f(\mathbf{x}) - \alpha \|\nabla f(\mathbf{x})\|^2 + R(-\alpha\nabla f(\mathbf{x})) < f(\mathbf{x})$$

$\forall\alpha < \bar{\alpha}$ $\mathbf{x}$ cannot be a local minimum. □

Proposition 1.1.2 states that the first order model allows to find the decreasing direction, but if the gradient is 0 we do not know if we are in presence of a minimum, maximum or saddle point. To discriminate among those, we exploit the information provided by the second derivative.

1.1.2 Second order model

Fact 1.1.3. *Let $f \in C^2$. If $\mathbf{x}$ is a local minimum then the gradient is positive semidefinite ($\nabla^2 f(\mathbf{x}) \succeq 0$).*

Proof by contraddiction. Our contradictory hypothesis is that we are in a local minimum, but the Hessian is not positive semidefinite (formally, $\exists d$ s.t. $d^T\nabla^2 f(\mathbf{x})d < 0$ or equivalently, $\exists\lambda_i < 0$, noticing that $\bar{f}(\alpha) = tr(\alpha H_i)\nabla f(\mathbf{x})(\alpha H_i) = \alpha^2\lambda_i < 0$).

Obs: saying that Hessian is not positive semidefinite means saying that there is a direction of negative curvature.

Just like in previous case, we take the direction d normalized ($\|d\| = 1$).

Let us consider a step $\mathbf{x}(\alpha) = \mathbf{x} + \alpha d$ and then take the second-order Taylor formula (since $\nabla f(x) = 0$ there is no linear term involved)

$$f(\mathbf{x}(\alpha)) = f(\mathbf{x}) + \frac{1}{2}\alpha^2 d^T\nabla^2 f(\mathbf{x})d + R(\alpha d)$$

with $\lim\limits_{\|h\|\to 0} \frac{R(h)}{\|h\|^2} = 0$, which means that the residual should go to 0 at least cubically.

Since $h = x - x(\alpha)$ we get that $\lim\limits_{\alpha\to 0} \frac{R(\alpha d)}{\alpha^2} = 0$ or equivalently $\forall\varepsilon > 0\, \exists\bar{\alpha} > 0$ s.t. $R(\alpha d) \leq \varepsilon\alpha^2 \forall 0 \leq \alpha < \bar{\alpha}$.

At this point, since this condition holds for each ϵ we are allowed to take the most convenient: $\varepsilon < -\frac{1}{2}d^T\nabla^2 f(x)d$, so that we obtain this condition on the residual $R(\alpha d) < -\frac{1}{2}\alpha^2 d^T\nabla^2 f(x)d$, hence

$$f(x(\alpha)) = f(x) + \frac{1}{2}\alpha^2 d^T\nabla^2 f(x)d + R(\alpha d) < f(x) \forall\, 0 \leq \alpha < \bar{\alpha}$$

Hence x cannot be a local minimum. □

In a local minimum, there cannot be directions of negative curvature "when the first derivative is 0, second-order effects prevail".

As far as sufficient conditions are concerned, we can prove the following

Fact 1.1.4. *Let $f \in C^2$ and let the Hessian be symmetric (hence real eigenvalues). If $\nabla f(x) = 0$ and the Hessian is strictly positive definite ($\nabla^2 f(x) \succ 0$) then x is a local minimum.*

Proof. Since the gradient is 0 we get the following second order Taylor approximation

$$f(x+d) = f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + R(d) \text{ with } \lim_{h \to 0} \frac{R(d)}{\|d\|^2} = 0$$

Hence, by definition of limit $\forall \varepsilon > 0 \ \exists \ \delta > 0$ s.t. $R(d) \leq \varepsilon \|d\|^2 \ \ \forall d$ s.t. $\|d\| < \delta$.

Since the Hessian is strictly positive definite $\lambda_{\min} > 0$ minimum eigenvalue of $\nabla^2 f(x)$, hence the variational caracterization of eigenvalues $d^T \nabla^2 f(x) d \geq \lambda_{\min} \|d\|^2$.

We are now ready to pick the ε we prefer ($\varepsilon < \lambda_{\min}$) to get $\forall d$ s.t. $\|d\| < \delta$

$$f(x+d) = f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + R(d) \geq f(x) + (\lambda_{\min} - \varepsilon) \|d\|^2 > f(x)$$

The term $\lambda_{\min} - \varepsilon$ is strictly positive □

In the remaining part of this lecture we will look for conditions that ensure that one a local minimum is found, it is also a global minimum.

Until now, we said that the local minima are those points where the fradient is 0 and the Hessian is positive semidefinite. An easy way to ensure that the Hessian is positive semidefinite in a ball around x is to have that the Hessian is positive semidefinite everywhere ($\forall x \in \mathbb{R}^n$) aka f is a convex function.

1.2 Convexity

Let us introduce some preliminaries to the hypothesis of convex functions.

Definition 1.2.1 (Convex hull). *Let $x, y \in \mathbb{R}^n$ we term **convex hull** and denote $conv(x, y) = \{z = \alpha x + (1-\alpha) y \ : \ \alpha \in [0,1]\}$ the segment joining x and y.*

Definition 1.2.2 (Convex set). *We term **convex set** if for each couple in the set, the line linksing such points belongs to the set.*

*Formally, $C \subset \mathbb{R}^n$ is a **convex set** if $\forall \ x, y \in C \ \ conv(x, y) \subseteq C$.*

Notice that "disconnected sets" cannot be convex sets.

Definition 1.2.3 (Convex hull of a set). *Given a set S, we can "complete" it to a convex set:*

$$\begin{aligned} conv(S) &= \bigcup \{ \, conv(x, y) \ : \ x, y \in S \, \} \\ &= \bigcap \{ \, C \ : \ C \text{ is convex } \wedge C \supseteq S \, \} \end{aligned}$$

Equivalently, the convex hull of S = iterated convex hull of all $x, y \in S$ or the smallest convex set containing S

Our goal is to find the nicest possible convex set that approximates our set.

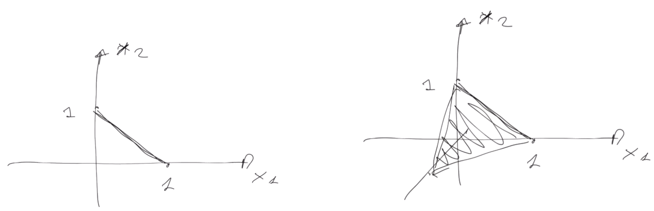
Fact 1.2.1. *A convex set is equal to its convex hull, formally C is convex $\iff C = conv(C)$.*

> **Note**
>
> A more general definition of a convex hull is the following: $\mathrm{conv}(\{x_1, \ldots, x_k\}) = \{x = \sum_{i=1}^k \alpha_i x_i \ : \ \sum_{i=1}^k \alpha_i = 1, \ \alpha_i \geq 0 \ \forall i\}$

Definition 1.2.4 (Unitary simplex)**.** *We term **unitary simplex** the set of k non-negative numbers summing to 1, formally*

$$\Theta^k = \{\alpha_i \in \mathbb{R}^k \ : \ \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0 \ \forall i\}$$

A few graphical examples are displayed in Figure 1.2.



(a) In $\mathbb{R}^2$ (b) In $\mathbb{R}^3$

FIGURE 1.2: Unitary simplexes.

We are interested in sufficient conditions for convexity.

Definition 1.2.5 (Cone)**.** *We term **cone** the set $\mathcal{C} = \{x \ : \ \alpha x \in \mathcal{C} \forall \alpha \geq 0\}$.*

An attentive reader may notice that the definition of cone is a relaxation of the unitary simplex, where we do not require the unitary sum.
The following sets are convex:

- Convex polytope $conv(\{x_1, \ldots, x_k\})$, unitary simplex Θ
- Affine hyperplane: $\mathcal{H} := \{x \in \mathbb{R}^n \ : \ ax = b\}$
- Affine subspace: $\mathcal{S} := \{x \in \mathbb{R}^n \ : \ ax \leq b\}$
- Ball in p-norm, $p \geq 1$: $\mathcal{B}_p(x, r) = \{y \in \mathbb{R}^n \ : \ \| y - x \|_p \leq r\}$
- Ellipsoid: $\mathcal{E}(Q, x, r) := \{y \in \mathbb{R}^n \ : \ (y - x)^T Q(y - x) \leq r\}$ with $Q \succeq 0$. Notice that ellipsoids are levelsets of quadratic functions.
- Open versions by substituting "<" to "≤"

- Cones
- Conical hull of a finite set of directions: $cone(\{d_1, \ldots, d_k\}) = \left\{ d = \sum_{i=1}^{k} \mu_i d_i \ : \ \mu_i \geq 0 \ \forall \, i \right\}$
- Lorentz (ice-cream) cone: $\mathbb{L} = \left\{ x \in \mathbb{R}^n \ : \ x_n \geq \sqrt{\sum_{i=1}^{n-1} x_i^2} \right\}$
- Cone of positive semidefinite matrices: $\mathbb{S}_+ = \{ A \in \mathbb{R}^{n \times n} \ : \ A \succeq 0\}$

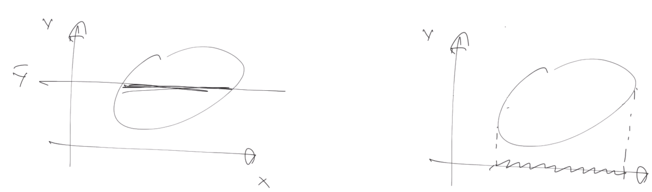
Fact 1.2.2. *The following operations preserve convexity.*

1. *Given a possibly infinite family of convex sets* ($\{C_i\}_{i \in I}$)*, the intersection* ($\bigcap_{i \in I} C_i$) *convex;*
2. *If we have convex sets in different subspaces, their cartesian product is a convex set* ($C_1, \ldots, C_k$ *convex* $\iff$ $C_1 \times \cdots \times C_k$ *convex);*
3. *Given a convex set, its image under a linear mapping (aka scaling, translation, rotation) is a convex set. Formally,* C *convex* $\Longrightarrow A(C) := \{x = Ay + b \ : \ y \in C\}$ *convex;*
4. C *convex* $\Longrightarrow A^{-1}(C) := \{x \ : \ Ax + b \in C\}$ *convex (inverse image under a linear mapping);*
5. *Let* C_1 *and* C_2 *convex and let* $\alpha_1, \alpha_2 \in \mathbb{R}$*. Then* $\alpha_1 C_1 + \alpha_2 C_2 := \{x = \alpha_1 x_1 + \alpha_2 x_2 \ : \ x_1 \in C_1, x_2 \in C_2\}$ *convex;*
6. $C \subseteq \mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ *convex* $\Longrightarrow$

   SLICE: $C(y) := \{x \in \mathbb{R}^{n_1} \ : \ (x, y) \in C\}$ *conve;*

   PROJECTION: $C^1 := \{x \in \mathbb{R}^{n_1} \ : \ \exists y \ s.t. \ (x, y) \in C\}$ *convex*

   A pictorial example in Figure 1.3;
7. C *convex* $\Longrightarrow int(C)$ *and* $cl(C)$ *convex*



(a) Slice on $\bar{y}$ (b) Project on the x component

FIGURE 1.3: Pictorial examples of slicing and projecting.

Theorem 1.2.3. *$\mathcal{P}$ is a polyhedron iff $\exists \{x_1, \ldots, x_k\}$ and $\{d_1, \ldots, d_h\}$ s.t. $\mathcal{P} = conv(\{x_1, \ldots, x_k\}) + cone(\{d_1, \ldots, d_h\})$.*

Notice that if we are interested in proving that a set with a certain shape is convex, we chould try to derive it from an object that we know is convex through the operations we enumerated above.

Definition 1.2.6 (Convex function)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function. We say that f is **convex** if $\forall x, y \in \mathbb{R}^n$, the segment that joins $f(x)$ and $f(y)$ lies above the function.*

*In other words, f is **convex** iff $epi(f)$ is convex, where epi denotes the epigraph of the function, graphically speaking, the region which is above the function line (in the plot).*

*Equivalently, we say that f is **convex** if $\forall x, y \in dom(f)$ for any $\alpha \in [0, 1]$, $\alpha f(x) + (1 - \alpha) f(y) \geq f(\alpha x + (1 - \alpha) y)$.*

Equivalently, $\forall x^1$, $\ldots$, x^k, $\alpha \in \Theta^k$

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

Definition 1.2.7 (Sublevel graph)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function. We term **sublevel graph** of $f(x)$ the projection on the x axis of the portions of the epigraph which lie below the constant $y = \bar{x}$.*

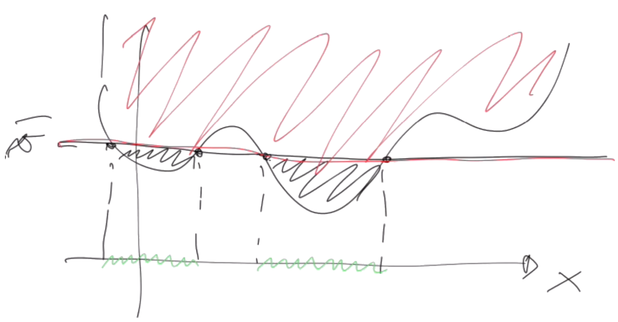


FIGURE 1.4: Pictorial example of sublevel graph. Such a graph is drawn in green in the figure.

Fact 1.2.4. *The following holds:*

- *Let f convex. Then $S(f, v)$ convex $\forall v \in \mathbb{R}$;*
- *f is concave if $-f$ is convex ("convex analysis is a one-sided world").*

The second statement of Proposition 1.2.4 is useful to make a comparison between minimizing and maximizing. In particular, if our aim is to maximize the function, we can be sure to have found a global maximum if the function is concave.

Definition 1.2.8 (Strict convexity)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We term f* ***strictly convex*** *iff $\alpha f(x) + (1-\alpha)f(y) > f(\alpha x + (1-\alpha)y)$.*

Definition 1.2.9 (Strong convexity)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We term f* ***strongly convex modulus*** *$\tau > 0$ iff $f(x) - \frac{\tau}{2} \|x\|^2$ is convex.*

Formally,

$$\alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y) + \frac{\tau}{2}\alpha(1-\alpha) \|(y-x)\|^2$$

Next lecture we will talk about how we can check tat a function is convex, operationally.