1 5th of October 2018 — A. Frangioni

1.1 Unconstrained optimization

Until now we stated that the best conditions 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 (Unconstrained optimization problem)**.** *We want to solve:*

$$f_* = \min\{f(x) \ : \ x \in X\} \ (P)$$

where $X = \mathbb{R}^n$*.*

If $\mathbb{R}^n$ is not bounded, Weierstrass theorem does not apply, hence even if a (global) minimum x_* exists, finding it is a NP problem.

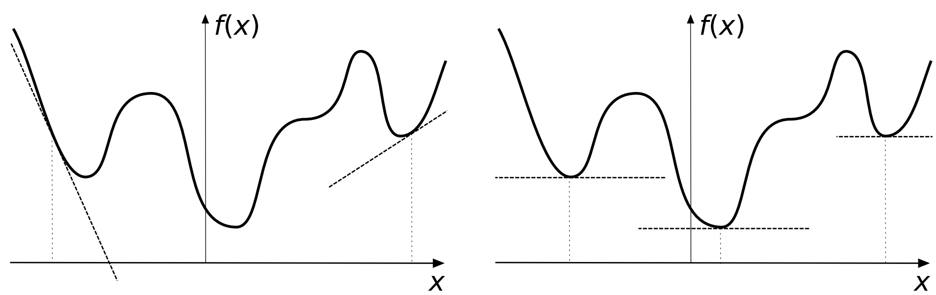
Let us use a weaker condition to ease things a little: x_* is a **local minimum** if it solves:

$$\min\{f(x) \ : \ x \in \mathcal{B}(x_*, \varepsilon)\}\text{for some } \varepsilon > 0$$

aka, the minimum we found is a global minimum in a ball around x^*.

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

To test these conditions derivatives help, as an example see Figure 1.1



(A) Not minima (B) Minima

FIGURE 1.1: In the leftmost plot, we can see that if the derivatives are non zero the point is not a minimum. Such a condition is satisfied in the right handed plot.

If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum.

Hence, $f'(x) = 0$ in all local minima, so this holds in the global one as well.

1.1.1 First order model

Do you recall?

The first order model of f is $L_x(y) = f(x) + \nabla f(x)(y-x)$, such that $f(y) = f(x) + \nabla f(x)(y-x) + R(y-x)$.
We already stated last lecture that if the norm of the argument of the residual is going to 0, then the residual is going to 0 faster (quadratically), formally $\lim\limits_{\|h\|\to 0} \frac{R(h)}{\|h\|} = 0$.

Fact 1.1. *Let f be differentiable, if x is a local minimum, then $\nabla f(x) = 0$.*

In which direction shall we move in order to get closer to the minimum, provided that we are sitting in x? $x(\alpha) = x - \alpha\nabla f(x)$, hence we should take a step along the anti-gradient $-\nabla f(x)$.

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

In our case, $y = x - \alpha\nabla f(x)$, so we get $f(x-\alpha\nabla f(x)) = f(x) - \alpha \|\nabla f(x)\|^2 + R(-\alpha\nabla f(x))$.

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

Take $\varepsilon < \|nabla f(x)\|$ to get $R(-\alpha\nabla f(x)) < \alpha \|\nabla f(x)\|^2$, then

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

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

Notice that the optimality condition also tells us how to move to get closer to the minimum.

An attentive reader may notice that the gradient is 0 in minima, maxima and saddle points (aka stationary point), hence how to discriminate among those?

We need to take infot account second derivatives, namely such second derivative should be positive for a minimum point.

1.1.2 Second order model

Fact 1.2. *Let $f \in C^2$. If x is a local minimum then $\nabla^2 f(x) \succeq 0$, in words the gradient is positive semidefinite.*

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

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 $x(\alpha) = x + \alpha d$ and then take the second-order Taylor formula (since $\nabla f(x) = 0$ there is no linear term involved)

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

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 1.3. *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 gradient 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 (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 1.3 (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 1.4 (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 1.4. *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 1.5 (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.

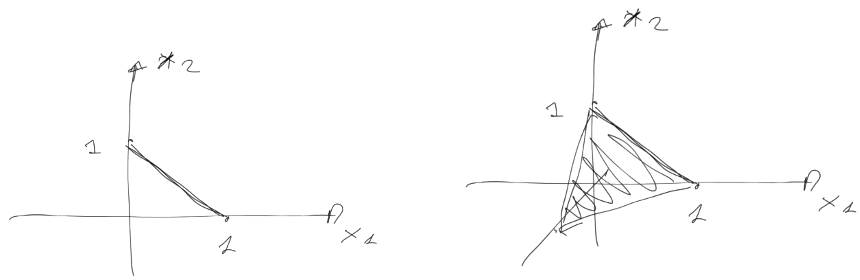
We are interested in sufficient conditions for convexity.

Definition 1.6 (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



(A) In $\mathbb{R}^2$ (B) In $\mathbb{R}^3$

FIGURE 1.2: Unitary simplexes.

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

Fact 1.5. *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;*

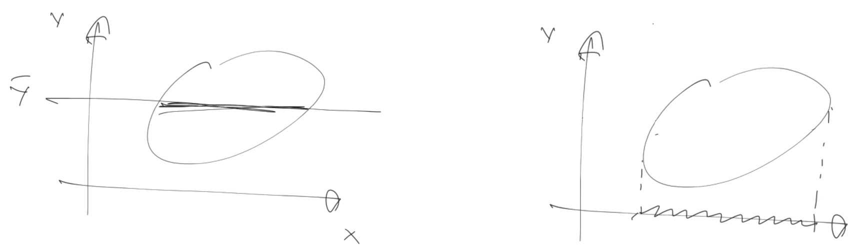
6. *$C \subseteq \mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ convex $\Longrightarrow$*

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

   PROJECTION: *$C^1 := \{x \in \mathbb{R}^{n_1} \ : \ \exists y \text{ s.t. } (x, y) \in C\}$ 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 1.6. *$\mathcal{P}$ 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 1.7 (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 1.8 (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 $y = \bar{x}$.*

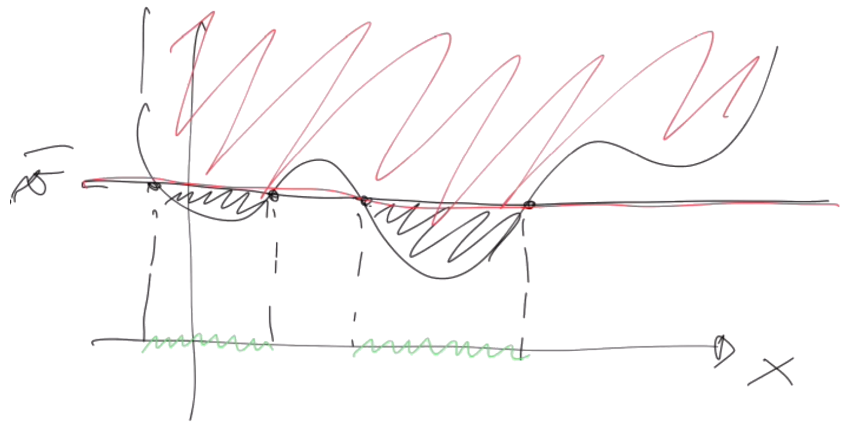


FIGURE 1.4: Pictorial example of sublevel graph. Such a graph is drawn in green in the figure.

Fact 1.7. *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 1.7 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 1.9 (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 1.10 (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.