Optimization for Data Science

F. Rinaldi[1]

[1] 
Università degli Studi di Padova


Padova
2020

Outline

Optimization for Data Science

Constrained Optimization

General Constrained Problem

We consider a problem of the form

$$\begin{aligned} &\min f(x) \\ &x \in C \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $C \subseteq \mathbb{R}^n$ is a convex set.

Useful Definitions

Definition [Feasible Directions]

Let $C \subseteq \mathbb{R}^n$ be a nonempty set. We define the set of *feasible directions* for C in $\bar{x} \in C$ the following set $F(\bar{x})$:

$$F(\bar{x}) = \{d \in \mathbb{R}^n,\ d \neq 0 :\ \exists\, \delta > 0 \text{ s.t. } \bar{x} + \alpha d \in C,\ \forall\, \alpha \in (0, \delta)\ \}.$$

Definition [Descent Directions]

We define *the set of descent directions* for f in $\bar{x}$ as follows:

$$D(\bar{x}) = \{d \in \mathbb{R}^n :\ \exists\, \delta > 0 \text{ such that } f(\bar{x} + \alpha d) < f(\bar{x}),\ \forall \alpha \in (0, \delta)\ \}.$$

Characterization of Local Minima

Proposition

Let $f \in C(\mathbb{R}^n)$. if $x^* \in C$ is a local (global) minimum of Problem (1) then

$$D(x^*) \cap F(x^*) = \emptyset \tag{2}$$

Some better Characterization

Proposition

Let $C \subseteq \mathbb{R}^n$ be a convex set and $\bar{x} \in C$ a feasible point for C. If $C \neq \{\bar{x}\}$, for all $x \in C$ with $x \neq \bar{x}$, direction

$$d = x - \bar{x}$$

is feasible for C in $\bar{x}$.

Definition of feasible directions for convex sets

$$F(\bar{x}) = \{d \in \mathbb{R}^n,\ d = x - \bar{x},\ x \in C,\ x \neq \bar{x}\}.$$

Necessary Conditions for Identifying Local Minima

Proposition [Necessary Conditions]

Let $x^* \in C$ be local minimum for problem

$$\begin{aligned} &\min f(x) \\ &x \in C \end{aligned}$$

with $C \subseteq \mathbb{R}^n$ convex and $f \in C^1(\mathbb{R}^n)$. Then

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall\, x \in C.$$

A Geometrical Representation of First Order Optimality Conditions

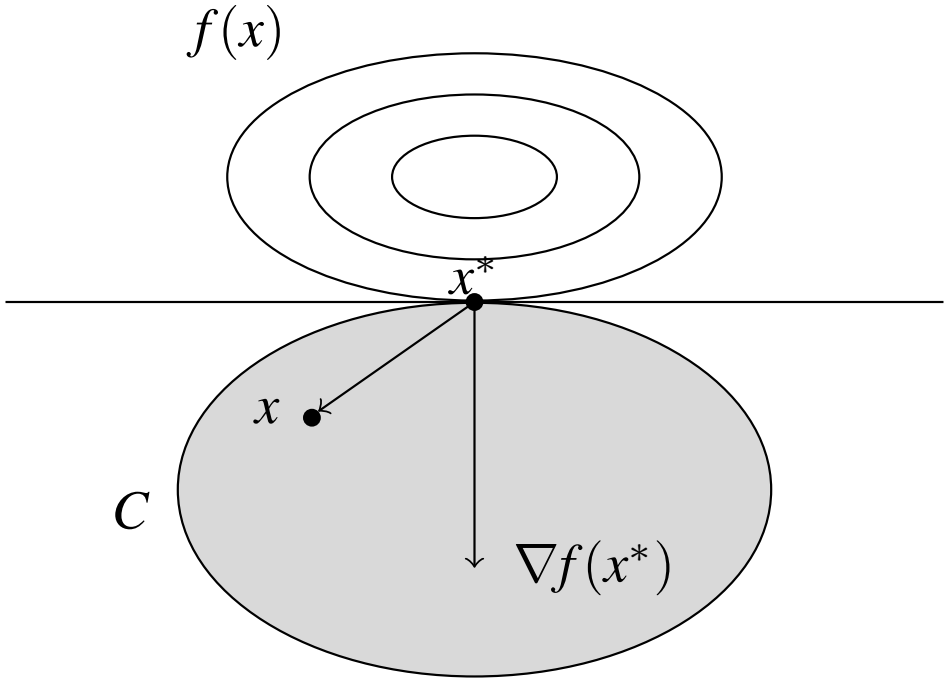

Figure: Geometrical representation of first order optimality condition.

Convex Case

Proposition [Necessary and Sufficient Conditions]

Let $C \subseteq \mathbb{R}^n$ be a convex set and $f \in C^1(\mathbb{R}^n)$ be a convex function. Point $x^* \in C$ is a global minimum of problem

$$\begin{aligned} &\min f(x) \\ &x \in C, \end{aligned}$$

if and only if

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall\, x \in C.$$

Remark

Strict convexity ensures uniqueness of the global minimum!

Projected Gradient Method

- **Gradient method** cannot be applied to constrained problems.
- Iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

might be such that

$$x_{k+1} \notin C.$$

- QUESTION: How to overcome the issue?
- ANSWER: Choose the point in C nearest to $x_k - \alpha_k \nabla f(x_k)$ as the new iterate.
- IDEA: Project the given point over the set C.

Projection

Definition

Let us consider:

- $\|\cdot\|$ euclidean norm;
- $C \subset \mathbb{R}^n$ a closed convex set;
- $\bar{x} \in \mathbb{R}^n$ a given point.

We define projection of $\bar{x}$ over C the solution $\rho_C(\bar{x})$ of the following problem:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - \bar{x}\|^2 \\ &x \in C. \end{aligned} \tag{3}$$

Basic Properties

Proposition [Projection Properties]

We have:

- $\bar{x} = \rho_C(\bar{x}), \quad \forall\, \bar{x} \in C;$
- $x^* \in C$ is projection of $\bar{x}$ over C, that is $x^* = \rho_C(\bar{x})$, if and only if

$$(\bar{x} - x^*)^\top (x - x^*) \leq 0 \quad \forall\, x \in C;$$

- projection operator is continuous and non-expansive:

$$\|\rho_C(y) - \rho_C(z)\| \leq \|y - z\| \quad \forall\, y, z \in \mathbb{R}^n.$$

Geometrical Representation of Projection Operator

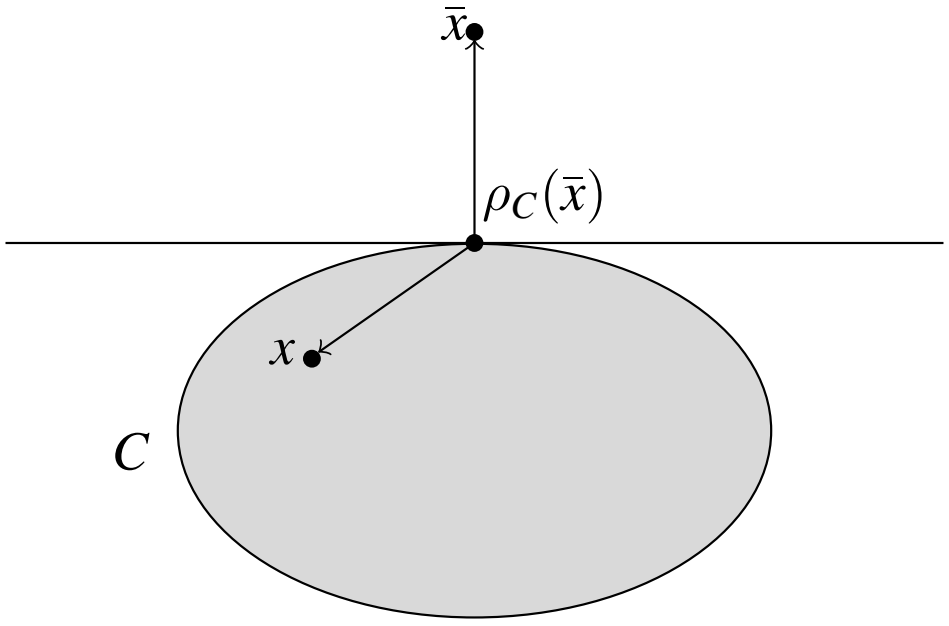

Figure: Geometrical representation of projection operator.

Projected Gradient Detailed Scheme

Algorithm 1 `Projected gradient method`

1 Choose a point $x_1 \in C$
2 For $k = 1, \ldots$
3 Set $\hat{x}_k = \rho_C(x_k - s_k \nabla f(x_k))$, with $s_k > 0$
4 If $\hat{x}_k$ satisfies some specific condition, then STOP
5 Set $x_{k+1} = x_k + \alpha_k(\hat{x}_k - x_k)$, with $\alpha_k \in (0, 1]$
suitably chosen stepsize
6 End for

Do we Get a Descent Direction?

- From properties of the projection, we have:

$$(x_k - s_k \nabla f(x_k) - \hat{x}_k)^\top (x - \hat{x}_k) \leq 0, \quad \forall\, x \in C.$$

- By setting $x = x_k$, we can write

$$(x_k - s_k \nabla f(x_k) - \hat{x}_k)^\top (x_k - \hat{x}_k) \leq 0$$

  and by properly rewriting, we get:

$$\nabla f(x_k)^\top (\hat{x}_k - x_k) \leq -\frac{1}{s_k} \|x_k - \hat{x}_k\|^2. \tag{4}$$

- d_k is a descent direction if $\|x_k - \hat{x}_k\| \neq 0$.

Iteration on the Projected Gradient

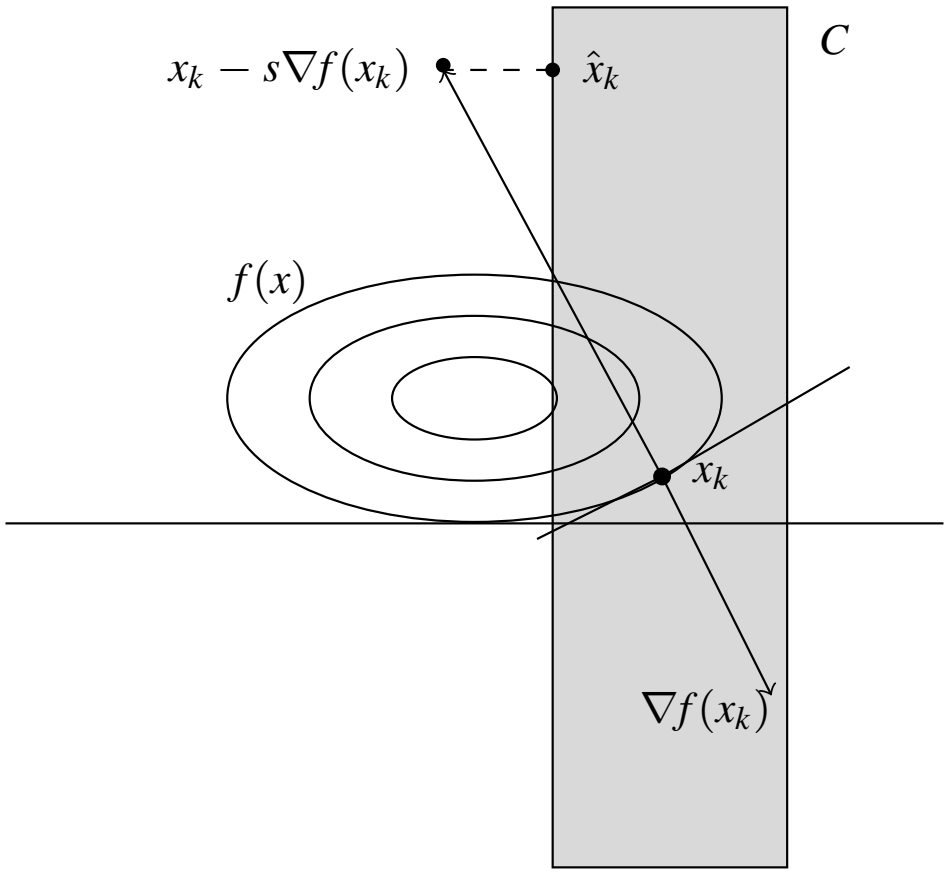

Figure: Iteration of the projected gradient.

Two versions of the Projected Gradient

Version 1 [Fixing s_k]

Fix s_k to constant value and use some line search to get $\alpha_k \in (0, 1]$.

- Line searches similar to the ones described in the unconstrained case.
- Main difference is that feasibility needs to be ensured, thus we get a stepsize $\alpha_k \in (0, 1]$.

Version 2 [Fixing α_k]

Fix α_k and implement a **curvilinear search** over s_k.

- Curvilinear search over C might be very costly.
- From now on, we only consider the case $s_k = s > 0$ constant.

Optimality Conditions using Projection

Proposition

Let $x^* \in C$ be local minimum of the problem:

$$\begin{aligned} &\min f(x) \\ &x \in C \end{aligned}$$

with $C \subseteq \mathbb{R}^n$ convex and $f \in C^1(\mathbb{R}^n)$. Then

$$x^* = \rho_C \left(x^* - s\nabla f(x^*) \right),$$

with $s > 0$.

- We obtain a condition that can be used to stop the algorithm at Step 3.

Proof

We first recall that

$$\hat{x}^* = \rho_C \left(x^* - s\nabla f(x^*) \right).$$

Taking into account previous results we get that

$$\nabla f(x^*)^\top (x - x^*) \geq 0,$$

for all $x \in C$.
From the properties of projection operator, we get

$$(x^* - s\nabla f(x^*) - \hat{x}^*)^\top (x^* - \hat{x}^*) \leq 0.$$

Combining the two inequalities, we get

$$(x^* - \hat{x}^*)^\top (x^* - \hat{x}^*) \leq -s\nabla f(x^*)^\top (\hat{x}^* - x^*) \leq 0.$$

That is

$$\|x^* - \hat{x}^*\| = 0,$$

and we have

$$x^* = \hat{x}^* = \rho_C \left(x^* - s\nabla f(x^*) \right).$$

□

Necessary and Sufficient Conditions: Convex Case

Proposition

Let $C \subseteq \mathbb{R}^n$ be a closed convex set and $f \in C^1(\mathbb{R}^n)$ be a convex function. Point $x^* \in C$ is a global minimum of the problem

$$\begin{aligned} &\min f(x) \\ &x \in C \end{aligned}$$

if and only if

$$x^* = \rho_C \left(x^* - s \nabla f(x^*) \right),$$

with $s > 0$.

Gradient Mapping

Definition [Gradient Mapping in x_k]

It is the vector

$$g_C(x_k) = x_k - \hat{x}_k,$$

where $\hat{x}_k = \rho_C(x_k - s_k \nabla f(x_k))$.

- We now fix $s = 1/L$ and $\alpha_k = 1/L$.
- Projection is $\hat{x}_k = \rho_C(x_k - \frac{1}{L}\nabla f(x_k))$.
- We then have that a generic iterate of the projected gradient method is

$$x_{k+1} = x_k - \frac{1}{L} g_C(x_k).$$

Properites of Gradient Mapping

Proposition

Let us consider a convex function with Lipschitz continuous gradient having constant $L > 0$. Then, we have

$$f(\hat{x}_k) - f(x_k) \leq -\frac{\|g_C(x_k)\|^2}{2L}.$$

If we further have that f is σ-strongly convex, we have

$$g_C(x)^\top (x - x^*) \geq \frac{\sigma}{2}\|x - x^*\|^2 + \frac{1}{2L}\|g_C(x)\|^2.$$

- First inequality very similar to decrease in Gradient Method!
- Second inequality needed to prove linear rate.

Rate of Projected Gradient

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function with Lipschitz continuous gradient having Lipschitz constant $L > 0$. Projected gradient method with fixed stepsize $\alpha_k = 1/L$ satisfies:

$$f(x_{k+1}) - f(x^*) \leq \frac{2L}{k}\|x_1 - x^*\|^2.$$

Furthermore, if f is σ-strongly convex, projected gradient method with fixed stepsize $\alpha_k = 1/L$ satisfies:

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma}{L}\right)^k \|x_1 - x^*\|^2.$$

Proof

[First part] Proof directly follows from analysis of the gradient method.

[Second part] Case when f is σ-strongly convex.

Taking into account the iteration k we can write

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \frac{1}{L} g_C(x_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - \frac{2}{L} g_C(x_k)^\top (x_k - x^*) + \frac{1}{L^2}\|g_C(x_k)\|^2 \\
&\leq \|x_k - x^*\|^2 - \frac{2}{L}\left[\frac{\sigma}{2}\|x_k - x^*\|^2 + \frac{1}{2L}\|g_C(x_k)\|^2\right] + \frac{\|g_C(x_k)\|^2}{L^2} \\
&= (1 - \frac{\sigma}{L})\|x_k - x^*\|^2.
\end{aligned}$$

By induction, we get

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma}{L}\right)^k \|x_1 - x^*\|^2.$$

□

Final Comments

PRO

Projected gradient guarantees similar convergence rates as the gradient method in the unconstrained case.

CON

For huge-scale problems computing the projection might be an expensive task!

- In some cases projection can be performed at a reasonable cost.

Cheap Projections

ℓ_2 ball: $C = \{x \in \mathbb{R}^n : \|\bar{x}\|_2 \leq 1\}$

In this case we have

$$\rho_C(\bar{x}) = \begin{cases} \bar{x} & \text{if } \|x\|_2 \leq 1 \\ \dfrac{\bar{x}}{\|\bar{x}\|_2} & \text{if } \|\bar{x}\|_2 > 1 \end{cases};$$

Box constraints: $C = \{x \in \mathbb{R}^n : l \leq x \leq u\}$

In this case we describe the projection component-wise

$$\rho_C(\bar{x})_i = \begin{cases} l_i & \text{if } \bar{x}_i < l_i \\ \bar{x}_i & \text{if } l_i \leq \bar{x}_i \leq u_i \\ u_i & \text{if } \bar{x}_i > u_i \end{cases},$$

for $i = 1, \ldots, n$.

- In both cases we have that projection costs $\mathcal{O}(n)$.