Universidade Federal do Piauí
Centro de Ciências da Natureza
Pós-Graduação em Matemática
Mestrado em Matemática

A subgradient method with non-monotone line search for Lipschitz convex functions

José Gonçalves de Oliveira Rufino

Teresina - 2024

José Gonçalves de Oliveira Rufino

Dissertação de Mestrado:

A subgradient method with non-monotone line search for Lipschitz convex functions

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Matemática, da Universidade Federal do Piauí, como requisito parcial para obtenção do grau de Mestre em Matemática.

Orientador:
Prof. Dr. João Carlos de Oliveira Souza

Teresina - 2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CENTRO DE CIÊNCIAS DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

A subgradient method with non-monotone line search for Lipschitz convex functions

José Gonçalves de Oliveira Rufino

Esta Dissertação foi submetida como parte dos requisitos necessários à obtenção do grau de **Mestre em Matemática**, outorgado pela Universidade Federal do Piauí.

Dissertação aprovada em 31 de julho de 2024.

Banca Examinadora:

João Carlos de Oliveira Sy
Prof. Dr. João Carlos de Oliveira Souza – Orientador

Jurandir de Oliveira Lopes
Prof. Dr. Jurandir de Oliveira Lopes – UFPI

Prof. Dr. Felipe Ignacio Lara Obreque – UTA

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

R926s Rufino, José Gonçalves de Oliveira.
A subgradient method with non-monotone line search for Lipschitz convex functions / José Gonçalves de Oliveira Rufino. -- 2024.
65 f. : il.

Dissertação (Mestrado) - Universidade Federal do Piauí, Centro de Ciências da Natureza, Programa de Pós-Graduação em Matemática, Teresina, 2024.
"Orientador: Prof. Dr. João Carlos de Oliveira Souza."

1. Método do gradiente. 2. Método do subgradiente. 3. Função convexa. 4. Método de gradiente não-monótono. I. Souza, João Carlos de Oliveira. II. Titulo.

CDD 515.64

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Dedicado aos meus avós: Maria Angélina Gonçalves Oliveira (Vó Gelina) e José Gonçalves de Oliveira, in memoriam, assim como a Eudóxia Custódio de Oliveira e Silva (Vó Dócinha) e José Rufino da Silva (Vô Zuca).

Agradecimentos

Agradeço a Deus pela oportunidade de estar vivo e pela força e coragem que me foram concedidas nas orações que fazia antes de ir para a universidade, em particular, na Ave Maria Sertaneja, cantada por minha prima Elda.

Agradeço ao meu orientador, professor Dr. João Carlos de Oliveira Souza, não só pelos planos, oportunidades e conselhos, que foram cruciais para minha formação acadêmica, mas também pela enorme disposição, paciência e amizade. O senhor é uma referência de professor e pesquisador para mim. Por meio do exemplo, é uma das grandes motivações que tenho para continuar estudando matemática. Também agradeço à Lara, que, juntamente com o senhor, foram as minhas pessoas aqui na cidade de Teresina. Muito obrigado.

Agradeço à minha família. Primeiro à minha mãe, Francisca Gonçalves, e ao meu pai, Raimundo Rufino, pela educação, apoio e amor que recebi. O poema de Manoel de Barros reflete um pouco do sentimento que tenho: "Acho que o quintal onde a gente brincou é maior do que a cidade. A gente só descobre isso depois de grande." A união desses fatores foi fundamental para que eu enfrentasse os desafios com a cabeça erguida. Além disso, meu pai é minha inspiração na matemática. Muito obrigado. Também agradeço aos meus irmãos, Emanuel e Raul, pelo apoio e incentivo.

Agradeço à Carla Roberta por todo o amor, compreensão e apoio que me foram concedidos, especialmente nos momentos mais difíceis. Eu te amo mil milhões, princesa.

Agradeço aos meus amigos: os de infância, do IFPI, do IGH, da graduação e do mestrado. Em particular, a Wilkreffy, a Gabriele e a Estevão. Vocês tornaram a jornada infinitamente mais divertida.

Agradeço aos professores Antônio Veloso, Fernando Barbosa e Valdemi Costa, cujas aulas

e conversas no IFPI me fascinaram e me incentivaram a estudar mais matemática e participar das olimpíadas.

Agradeço aos membros da banca, professores Dr. Jurandir de Oliveira Lopes e Dr. Felipe Ignacio Lara Obreque, por terem participado e por seus comentários e sugestões, que ajudaram a melhorar a versão final da dissertação.

Agradeço aos professores da UFPI que ajudaram a iluminar este caminho.

Agradeço à CAPES pelo apoio financeiro.

"*And as we wind on down the road*
Our shadows taller than our soul
There walks a lady we all know
Who shines white light and wants to show
How everything still turns to gold
And if you listen very hard
The tune will come to you, at last
When all are one and one is all, yeah
To be a rock and not to roll
And she's buying a stairway to heaven."

Stairway to heaven - by Heart.

Resumo

Neste trabalho, estudamos resultados de convergência dos métodos clássicos do gradiente e do subgradiente, além de uma variação do método subgradiente com busca linear não monótona para funções convexas Lipschitz. O método do gradiente é um método de descida e os tamanhos de passo são escolhidos de forma exata e inexata com busca linear. O método subgradiente não é necessariamente um método de descida e os tamanhos de passo estudados são pré-fixados, não sendo escolhidos via busca linear. Assim, também estudamos um método subgradiente com busca linear não monótona que, apesar de não ser um método de descida, o possível aumento nos valores da função é controlado e os tamanhos de passo são escolhidos de forma adaptativa.

Palavras-chave: Método do gradiente, método do subgradiente, método do subgradiente não-monótono, função convexa.

Abstract

In this work, we study convergence results of the classical gradient and subgradient methods, as well as a variation of the subgradient method with non-monotone line search for convex Lipschitz functions. The gradient method is a descent method and step sizes are chosen exactly and inexactly with line search. The subgradient method is not necessarily a descent method and the step sizes studied are pre-fixed and are not chosen via line search. Thus, we also studied a subgradient method with non-monotone line search which, despite not being a descent method, the possible increase in function values is controlled and step sizes are chosen adaptively.

Keywords: Gradient method, subgradient method, non-monotone subgradient method, convex function.

List of Figures

Contents

Introduction

In this work, we study the subgradient method with non-monotone line search for Lipschitz convex functions proposed in [8]. The method performs a line search similar to the Armijo line search commonly used in the differentiable context. Additionally, we provide a brief study of classical gradient and subgradient methods along with their respective characteristics.

The gradient method, also known as the Cauchy method, is one of the oldest and most well-known strategies for minimizing a multivariable function. Its theoretical simplicity makes it particularly applicable to high-dimensional problems. However, computationally, the gradient method can exhibit a "zig-zag" behavior, resulting in slower convergence. Nevertheless, it serves as a fundamental basis for the development and refinement of more efficient methods.

The gradient method in each iteration moves in the direction opposite to the gradient vector, with a certain step size to ensure the descent algorithm. The proper choice of step size plays a crucial role in the method's effectiveness. Common approaches include the fixed step size, the one-dimensional minimization rule (exact line search), and the Armijo rule (inexact line search). Here are some basic references on the gradient method and its convergence properties: [2, 3, 11, 15, 17, 18].

The subgradient method for solving nondifferentiable convex optimization problems was developed in the 1960s, as evidenced by [7, 21]. In each iteration, a step is taken in the direction opposite to a subgradient. It is not necessarily a descent method because the direction opposite to a subgradient may not be a descent direction.

In the classical case, the sequence of step sizes is predetermined before the algorithm starts, and the step sizes are not chosen via line search. Five typical choices are considered, where the sequence of step sizes is either constant or tends to zero at a sublinear rate: constant step size, constant step length, square summable but not summable, nonsummable

diminishing, and nonsummable diminishing step lengths. Here are some fundamental references on the subgradient method and its convergence properties: [2, 11, 16, 19, 21].

The subgradient method proposed in [8] for minimizing Lipschitz convex functions includes a line search performed in the direction opposite to a subgradient. This search allows the function to increase over iterations, but the increase is controlled by a sequence (of nonmonotonicity) of nonincreasing parameters. The method utilizes this nonmonotone search to determine the step size, implying that this variant of the subgradient method has adaptive step sizes. Furthermore, since the search depends on the nonmonotonicity sequence, the step size is implicitly controlled by it.

This work is structured as follows: In Chapter 1, we present some notations, definitions, and results from optimization theory that will be used throughout the work. In Chapter 2, we study descent methods and some common line search techniques in the differentiable case. We introduce the gradient method and convergence results under certain assumptions on the function , its gradient, and choices of search rules. In Chapter 3, we highlight the differences that arise in the study of convex optimization when moving from the differentiable to the nondifferentiable context. We present the subgradient method, step size rules, classical inequalities, and convergence results. In Chapter 4, we introduce the algorithm "subgrad projection method with non-monotone line search", study some inequalities, and present convergence results of the method under assumptions on the nonmonotonicity sequence.

Chapter 1

Preliminaries

In this chapter, we present some notations, definitions, and results in optimization theory that will be used throughout the work, which can be found in [1, 5, 10, 11, 22].

Let $D \subset \mathbb{R}^n$ and $\Omega \subset \mathbb{R}^n$ such that $D \subset \Omega$, and let $f : \Omega \to \mathbb{R}$ be a function. Consider the problem of minimizing f over the set D, that is,

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in D. \end{aligned} \tag{1.1}$$

The set D is referred to as the feasible set of the problem, the points in D are called feasible points, and f is referred to as the objective function.

Definition 1.0.1. *We say that a point $\bar{x} \in D$ is*

1. *a global minimizer of (1.1) if*

$$f(\bar{x}) \leq f(x), \quad \forall x \in D;$$

2. *a local minimizer of (1.1) if there exists a neighborhood U of $\bar{x}$ such that*

$$f(\bar{x}) \leq f(x), \quad \forall x \in D \cap U.$$

Proposition 1.0.1. *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at the point $\bar{x}$. If $\bar{x}$ is an unconstrained local minimizer of f, then*

$$\nabla f(\bar{x}) = 0.$$

Definition 1.0.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L_{f,\mathcal{C}}$-Lipschitz continuous on $\mathcal{C} \subset \mathbb{R}^n$ if there exist a constant $L_{f,\mathcal{C}} > 0$ such that $|f(x) - f(y)| \leq L_{f,\mathcal{C}}\|x - y\|$, for all $x, y \in \mathcal{C}$. Whenever $C = \mathbb{R}^n$ we set $L_f \equiv L_{f,\mathbb{R}^n}$.*

Proposition 1.0.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on $\mathbb{R}^n$, with Lipschitz continuous gradient in $\mathbb{R}^n$ with constant $L > 0$. Then,*

$$|f(x+y) - f(x) - \langle \nabla f(x), y\rangle| \leq \frac{L}{2}\|y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

Definition 1.0.3. *A set $A \subset \mathbb{R}^n$ is said to be convex if, for every $x, y \in A$ we have*

$$\lambda x + (1-\lambda)y \in A, \quad \forall \lambda \in [0,1].$$

Proposition 1.0.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function on a convex and open set $\Omega \subset \mathbb{R}^n$. Then, for every $x, y \in \Omega$ there exists $t \in [0,1]$ such that*

$$f(y) - f(x) = \langle \nabla f(tx + (1-t)y), y - x\rangle.$$

Definition 1.0.4. *A function $f : A \subset \mathbb{R}^n \to \mathbb{R}$ is said to be convex if A is convex and for every $x, y \in A$, the inequality holds:*

$$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall \lambda \in [0,1].$$

Proposition 1.0.4. *Let $\Omega \subset \mathbb{R}^n$ be a convex and open set, and $f : \Omega \to \mathbb{R}$ a differentiable function on Ω. If f is convex on Ω, then for every $x, y \in \Omega$, we have*

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

Definition 1.0.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We say that $y \in \mathbb{R}^n$ is a subgradient of f at the point $x \in \mathbb{R}^n$ if*

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

The set of all subgradients of f at x is called the subdifferential of f at x denoted by $\partial f(x)$.

Proposition 1.0.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then for every $x \in \mathbb{R}^n$, the set $\partial f(x)$ is convex, compact, and non-empty. Moreover, for every $d \in \mathbb{R}^n$, we have*

$$f'(x; d) = \max\{\langle y, d\rangle \mid y \in \partial f(x)\}.$$

Proposition 1.0.6. *A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at the point $x \in \mathbb{R}^n$ if and only if the set $\partial f(x)$ contains exactly one element. In this case, $\partial f(x) = \{\nabla f(x)\}$.*

Definition 1.0.6. *Let $D \subset \mathbb{R}^n$ be a convex set and $\bar{x} \in D$. The normal cone at the point $\bar{x}$ with respect to the set D is defined as*

$$\mathcal{N}_D(\bar{x}) = \{d \in \mathbb{R}^n \mid \langle d, x - \bar{x}\rangle \leq 0, \quad \forall x \in D\}.$$

Proposition 1.0.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $D \subset \mathbb{R}^n$ a convex set. Then $\bar{x} \in \mathbb{R}^n$ is a minimizer of f on D if and only if*

$$\exists\, y \in \partial f(\bar{x}) \text{ such that } \langle y, x - \bar{x}\rangle \geq 0, \quad \forall x \in D,$$

or equivalently,

$$0 \in \partial f(\bar{x}) + \mathcal{N}_D(\bar{x}).$$

In particular, $\bar{x}$ is a minimizer of f in $\mathbb{R}^n$ if and only if

$$0 \in \partial f(\bar{x}).$$

Proposition 1.0.8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the set where the function f is not differentiable has Lebesgue measure zero.*

Definition 1.0.7. *A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is said to be σ-strongly convex with modulus $\sigma \geq 0$ if $f(\tau x + (1-\tau)y) \leq \tau f(x) + (1-\tau)f(y) - \frac{\sigma}{2}\tau(1-\tau)\|x-y\|^2$, for all $x, y \in \mathbb{R}^n$ and $\tau \in [0,1]$.*

Proposition 1.0.9. *The function $f : \mathbb{R}^n \to \mathbb{R}$ is σ-strongly convex with modulus $\sigma \geq 0$ if and only if $f(y) \geq f(x) + \langle v, y - x\rangle + (\sigma/2)\|y-x\|^2$, for all $x, y \in \mathbb{R}^n$ and all $v \in \partial f(x)$.*

Proposition 1.0.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex. Then, for all $x \in \mathbb{R}^n$ the set $\partial f(x)$ is a non-empty, convex, compact subset of $\mathbb{R}^n$. In addition, f is $L_{f,\mathcal{C}}$-Lipschitz function on $\mathcal{C} \subset \mathbb{R}^n$ if and only if $\|v\| \leq L_{f,\mathcal{C}}$ for all $v \in \partial f(x)$ and $x \in \mathcal{C}$.*

Definition 1.0.8. *Let $\mathcal{C} \subset \mathbb{R}^n$ be a non-empty, closed and convex set. The projection map, denoted by $\mathcal{P}_{\mathcal{C}} : \mathbb{R}^n \rightrightarrows \mathcal{C}$, is defined as follows $\mathcal{P}_{\mathcal{C}}(y) := \arg\min\{\|y - z\| : \ z \in \mathcal{C}\}$.*

Proposition 1.0.11. *Let $D \subset \mathbb{R}^n$ be a non-empty, convex and closed set. Then for every $x \in \mathbb{R}^n$, the projection of x onto D, $\mathcal{P}_D(x)$, exists and is unique.*

Furthermore, $\bar{x} = \mathcal{P}_D(x)$ if and only if

$$\bar{x} \in D, \quad \langle x - \bar{x}, y - \bar{x}\rangle \leq 0, \quad \forall y \in D,$$

or equivalently,

$$\bar{x} \in D, \quad x - \bar{x} \in \mathcal{N}_D(\bar{x}).$$

Proposition 1.0.12. *Le $D \subset \mathbb{R}^n$ be a non-empty, convex and closed set. Then, for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we have*

$$\langle \mathcal{P}_D(x) - \mathcal{P}_D(y), x - y \rangle \geq ||\mathcal{P}_D(x) - \mathcal{P}_D(y)||^2 \geq 0,$$

and

$$||\mathcal{P}_D(x) - \mathcal{P}_D(y)|| \leq ||x - y||.$$

In particular, $\mathcal{P}_D(\cdot)$ is continuous on $\mathbb{R}^n$.

Proposition 1.0.13. *Let $y \in \mathbb{R}^n$ and $z \in \mathcal{C}$. Then, we have $\|\mathcal{P}_{\mathcal{C}}(y) - z\|^2 \leq \|y - z\|^2$.*

Definition 1.0.9. *Let S be a nonempty subset of $\mathbb{R}^n$. A sequence $(v^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is said to be quasi-Fejér convergent to S, if and only if, for all $v \in S$ there exists $\bar{k} \geq 0$ and a summable sequence $(\epsilon_k)_{k\in\mathbb{N}}$, such that $\|v^{k+1} - v\|^2 \leq \|v^k - v\|^2 + \epsilon_k$ for all $k \geq \bar{k}$.*

Proposition 1.0.14. *Let $(v^k)_{k\in\mathbb{N}}$ be quasi-Fejér convergent to S. Then, the following conditions hold:*

(i) the sequence $(v^k)_{k\in\mathbb{N}}$ is bounded;

(ii) if a cluster point $\bar{v}$ of $(v^k)_{k\in\mathbb{N}}$ belongs to S, then $(v^k)_{k\in\mathbb{N}}$ converges to $\bar{v}$.

Chapter 2

Gradient method

The chapter will be divided into two sections. In the first section, we will present the idea of descent methods and line search techniques, which are natural in the differentiable case. The main results are the descent lemma and the rules for line search. We will use the descent lemma as a criterion for choosing a descent direction and the rules for line search as a criterion for choosing the step size. In the second section, we will present the gradient method, which is a descent method and in each iteration performs a search in the direction opposite to the gradient vector. The main results are the convergence theorems, under certain assumptions on the function f, its gradient f and the choice of search rules.

The functions considered are once or twice differentiable.

The construction of the first and second sections were based on references [2], [3], [11], [15], [17], [18], [20].

2.1 Descent methods. Line search.

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and the unconstrained minization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n. \end{aligned} \tag{2.1}$$

An idea to solve this problem is as follows: Starting from a point $x^k \in \mathbb{R}^n$, we want to find a new point $x^{k+1} \in \mathbb{R}^n$ such that

$$f(x^{k+1}) < f(x^k).$$

This new point can be obtained, from the point x^k, taking a direction $d^k \in \mathbb{R}^n$ according to which f decreases, at least for steps sufficiently small, and choosing a step size $t_k > 0$ such that

$$f(x^k + t_k d^k) < f(x^k).$$

Thus, we take $x^{k+1} := x^k + t_k d^k$ and repeat the procedure for the new point obtained. In this way, we construct a sequence $\{x^k\}$ with the property that $f(x^{k+1}) < f(x^k)$, for each $k = 0, 1, 2, ...$ Methods that use this idea are called descent methods.

In the idea about how to obtain the point x^{k+1} there are no instructions on how to choose a descent direction d^k and a step size t_k. In general, there are several possible descent directions and step size to take.

Definition 2.1.1. *We say that $d \in \mathbb{R}^n$ is a descent direction of the function $f : \mathbb{R}^n \to \mathbb{R}$ at the point $x \in \mathbb{R}^n$ if there exists $\varepsilon > 0$ such that*

$$f(x + td) < f(x) \quad \forall t \in (0, \varepsilon].$$

We denote by $\mathcal{D}_f(x)$ the set of all descent directions of the function f at the point x.

Of course, $\mathcal{D}_f(x)$ can be empty, for example, if x is a minimizer of the Problem (2.1); and it is clear that $\mathcal{D}_f(x) \cup \{0\}$ is a cone, since if $d \in \mathcal{D}_f(x)$, for all $t > 0$ we have that $td \in \mathcal{D}_f(x)$ and if $t = 0$ then $td \in \{0\}$.

From this definition, we mathematically formalize the meaning of $d \in \mathbb{R}^n$ being a direction according to which f decreases, at least for sufficiently short steps. In other words: to show, by definition, that d is a descent direction, we have to guarantee that there exists $\varepsilon > 0$ such that for all values of $t \in (0, \varepsilon]$, the inequality $f(x + td) < f(x)$ is true.

If we add the hypothesis that f is differentiable at the point x, then it is possible to obtain a more practical characterization for deciding whether d is a descent direction of f at the point x.

Lemma 2.1.1. *Consider $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function at the point $x \in \mathbb{R}^n$. Then:*

a) For all $d \in \mathcal{D}_f(x)$, we have $\langle \nabla f(x), d \rangle \leq 0$.

b) If $d \in \mathbb{R}^n$ satisfies $\langle \nabla f(x), d \rangle < 0$ then $d \in \mathcal{D}_f(x)$.

Proof. Let $d \in \mathcal{D}_f(x)$. Then there exists $\varepsilon > 0$ such that $f(x+td) < f(x)$, for all $t \in (0, \varepsilon]$. Since f is differentiable at the point x, for all $t > 0$, $f(x+td) = f(x) + \langle \nabla f(x), td \rangle + o(t)$, with $\lim_{t \to 0_+} \frac{o(t)}{t} = 0$. So,

$$0 > f(x+td) - f(x) = t\left(\langle \nabla f(x), d \rangle + \frac{o(t)}{t} \right), \quad \forall t \in (0, \varepsilon].$$

Dividing both sides of the above inequality by $t > 0$ and taking the limit as $t \to 0_+$, we obtain that $0 \geq \langle \nabla f(x), d \rangle$, which shows item (a).

Now suppose that $\langle \nabla f(x), d \rangle < 0$. Again, since f is differentiable at the point x, for $t > 0$, we have

$$f(x+td) - f(x) = t\left(\langle \nabla f(x), d \rangle + \frac{o(t)}{t} \right),$$

with $\lim_{t \to 0_+} \frac{o(t)}{t} = 0$. As $\lim_{t \to 0_+} \frac{o(t)}{t} = 0$ and $0 < -\frac{1}{2}\langle \nabla f(x), d \rangle$, then there exists a δ such that if $t \in (0, \delta)$ we have that

$$\frac{o(t)}{t} \leq -\frac{1}{2}\langle \nabla f(x), d \rangle.$$

Thus,

$$\begin{aligned} \langle \nabla f(x), d \rangle + \frac{o(t)}{t} &\leq \langle \nabla f(x), d \rangle - \frac{1}{2}\langle \nabla f(x), d \rangle \\ &= \frac{1}{2}\langle \nabla f(x), d \rangle \\ &< 0. \end{aligned}$$

Therefore,

$$\begin{aligned} f(x+td) - f(x) &= t\left(\langle \nabla f(x), d \rangle + \frac{o(t)}{t} \right) \\ &\leq t\left(\frac{1}{2}\langle \nabla f(x), d \rangle \right) \\ &< 0, \quad \forall\, t \in (0, \delta), \end{aligned}$$

that is, $d \in \mathcal{D}_f(x)$, which shows part (b). □

In the differentiable context, the lemma above is a tool to choose d as a direction of descent of f at the point x, for example, we can prove that, if $\nabla f(x) \neq 0$, then $-\nabla f(x) \in \mathcal{D}_f(x)$. In fact, taking $d = -\nabla f(x)$ we have that

$$\langle \nabla f(x), d \rangle = \langle \nabla f(x), -\nabla f(x) \rangle = -\|\nabla f(x)\|^2 < 0.$$

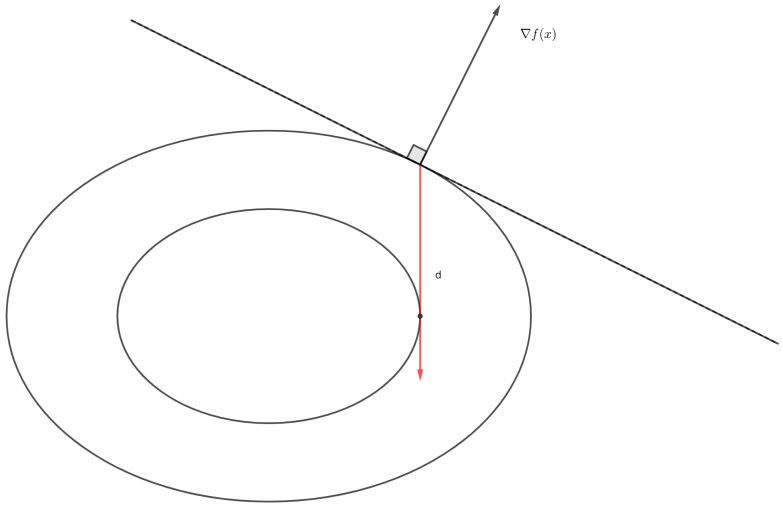


Figure 2.1: Illustration of item b) of the Lemma 2.1.1.

The Lemma 2.1.1 shows that, for $d^k \in \mathbb{R}^n$ to be a descent direction, it is sufficient that $\langle \nabla f(x), d^k \rangle < 0$, see Figure 2.1. Then let us consider the following descent algorithm to solve the Problem (2.1).

Algorithm 1 Descent algorithm

1: Choose $x^0 \in \mathbb{R}^n$ and set $k := 0$;
2: **while** $||\nabla f(x^k)|| \neq 0$ **do**
3: Choose $d^k \in \mathbb{R}^n$ such that $\langle \nabla f(x^k), d^k \rangle < 0$;
4: Choose $t_k > 0$ such that $f(x^k + t_k d^k) < f(x^k)$;
5: Set $x^{k+1} := x^k + t_k d^k$ and update $k := k + 1$;
6: **end while**

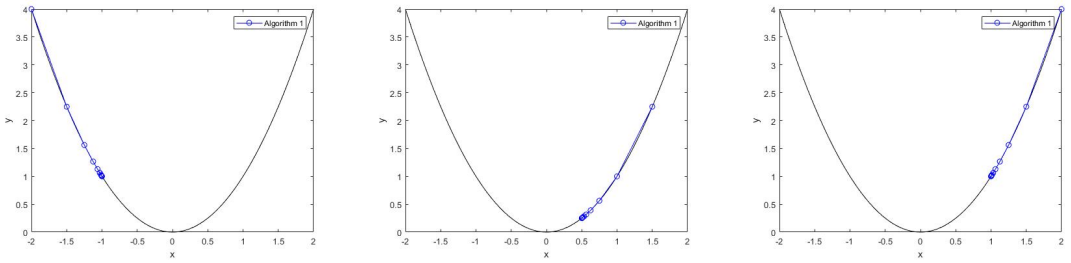
Since $\nabla f(x^k) \neq 0$, we have already seen that it is possible to perform step 3. The Lemma 2.1.1 guarantees that the direction d^k chosen belongs to the set $\mathcal{D}_f(x^k)$, which means, by the very definition of $\mathcal{D}_f(x^k)$, that it is possible to take $t_k > 0$ such that $f(x^k + t_k d^k) < f(x^k)$, therefore it is possible to execute the step 4. As we already have the point x^k and we have already chosen d^k and t_k, it is possible to define x^{k+1} with the desired descent property and update the value of k to perform the procedure for the new point x^{k+1} obtained, that is, it is possible to execute step 5.

By the structure of Algorithm 1, we either find a critical point after a finite number of iterations or generate a sequence $\{x^k\}$ such that the sequence $\{f(x^k)\}$ is decreasing.

Let $\{x^k\}$ be a sequence generated by the Algorithm 1. The next examples show that

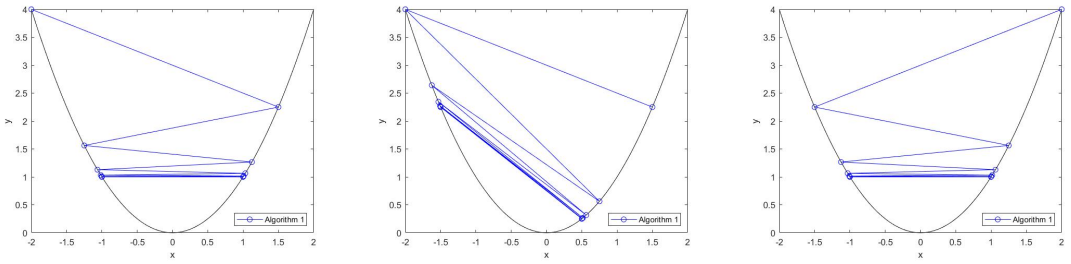
$\{x^k\}$ can have cluster points that are not critical points of the Problem (2.1). Furthermore, they show that the choice of step size t_k influences the convergence of the sequence $\{x^k\}$. If the step size is too small, the sequence $\{x^k\}$ may converge to a point that is not critical and if the step size is too large, the sequence $\{x^k\}$ may not converge and generate non-critical cluster points.

Example 2.1.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. Clearly, f has a critical point (global minimum) at $x^* = 0$. Note that $d = -1 \in \mathcal{D}_f(x)$, for every $x > 0$ and $d = 1 \in \mathcal{D}_f(x)$, for every $x < 0$. We run Algorithm 1 for different starting points with the stepsize $t_k = \frac{1}{2^{k+1}}$ and $t_k = 2 + \frac{3}{2^{k+1}}$ with the stop rule while $k \leq 500$.*



(a) $x^0 = -2$ (b) $x^0 = 1.5$ (c) $x^0 = 2$

Figure 2.2: Algorithm 1 for Example 2.1.1 with $t_k = \frac{1}{2^{k+1}}$.



(a) $x^0 = -2$ (b) $x^0 = 1.5$ (c) $x^0 = 2$

Figure 2.3: Algorithm 1 for Example 2.1.1 with $t_k = 2 + \frac{3}{2^{k+1}}$.

In the Figure 2.2, since we chose $t_k = \frac{1}{2^{k+1}}$ then t_k converges to 0 when k goes to infinity. For example, when $k = 50$, $t_{50} = \frac{1}{2^{50+1}} = 4.4409e^{-16}$; when $k = 200$, $t_{200} = \frac{1}{2^{200+1}} = 3.1115e^{-61}$; when $k = 500$, $t_{500} = \frac{1}{2^{500+1}} = 1.5275e^{-151}$. This shows that over the first few iterations the step size becomes so small that the sequence $\{x_k\}$ converges on a point that is not critical.

In Figure 2.3, since we chose $t_k = 2 + \frac{3}{2^{k+1}}$ then t_k converges to 2 when k goes to infinity. For example, when $k = 50$, $t_{50} = 2 + \frac{3}{2^{50+1}} = 2 + 1.3323e^{-15}$; when $k = 200$, $t_{200} = 2 + \frac{3}{2^{200+1}} = 2 + 9.3345e^{-61}$; when $k = 500$, $t_{500} = 2 + \frac{3}{2^{500+1}} = 2 + 4.5824e^{-151}$. This shows that during the first few iterations the step size is already close to 2. Since 2 is a large step size, the sequence $\{x_k\}$ skips the critical point and generates two cluster points that are not critical.

A good choice of step size t_k consists of preventing the step length from being too small or too large, and balancing this with the decrease promoted in the function f.

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, a point x^k and a direction $d^k \in \mathcal{D}_f(x^k)$, a natural strategy is to search for a suitable t_k along the half-line $x^k + td^k$, $t \geq 0$.

Line search. The step size is calculated by observing the behavior of the function f along the half-line $x^k + td^k$, $t \geq 0$, or along a limited interval in the same direction.

Among the line search rules, we will look at the one-dimensional minimization rule (exact search), Armijo rule (inexact search) and the fixed step size rule. Let us fix $x^k \in \mathbb{R}^n$ and $d^k \in \mathcal{D}_f(x^k)$.

One-dimensional minimization rule. The strategy is to minimize the objective function on the half-line $x^k + td^k$, $t \geq 0$. The step size t_k is given by the condition

$$f(x^k + t_k d^k) = \min_{t \geq 0} f(x^k + td^k),$$

i.e,

$$t_k = \arg\min_{t \geq 0} f(x^k + td^k). \tag{2.2}$$

Since we chose $d^k \in \mathcal{D}_f(x^k)$, there exists $\varepsilon > 0$ such that $f(x^k + td^k) < f(x^k)$ for all $t \in (0, \varepsilon]$. By the structure of the rule, this t_k is the "best" possible in the sense that for all $t \geq 0$, $f(x^k + t_k d^k) \leq f(x^k + td^k)$.

An advantage of using the one-dimensional minimization rule is that, in each iteration of the method, we choose the "best" t_k possible, i.e. the t_k that decreases the function f the most along the corresponding half-line. One disadvantage is that, in each iteration, it is necessary to solve a one-dimensional minimization Subproblem (2.2) in order to choose the value of t_k.

Remark 2.1.1. *By relation (2.2), we have that* $t_k = arg\min_{t \geq 0} \varphi_k(t)$, *where* $\varphi_k : \mathbb{R}_+ \to \mathbb{R}$, $\varphi_k(t) = f(x^k + td^k)$. *Since* $d^k \in \mathcal{D}_f(x^k)$ *then* $f(x^k + t_k d^k) < f(x^k + 0d^k)$, *this guarantees us that* $t_k > 0$. *In this case, if the function* f *is differentiable at the point* x^{k+1}, *by the*

Proposition 1.0.1 we have that

$$0 = \varphi_k'(t_k) = \langle \nabla f(x^k + t_k d^k), d^k \rangle = \langle \nabla f(x^{k+1}), d^k \rangle. \tag{2.3}$$

See the Figure 2.1 below

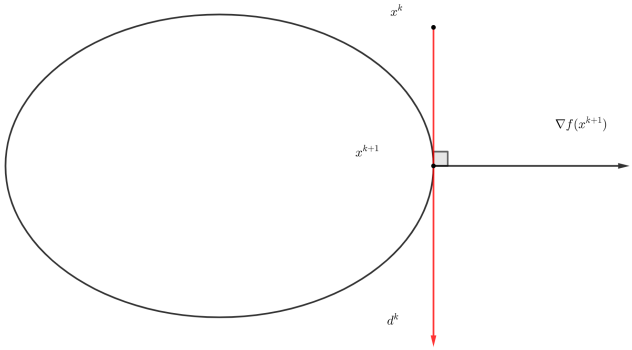


Figure 2.4: Illustration of the property 2.3

Armijo rule. The idea is to find a step size α_k that provides a reasonable decrease in the function f, without trying to minimize it. Suppose that f is differentiable at the point x^k. We fix the parameters $\hat{\alpha} > 0$, σ, $\theta \in (0, 1)$. We take $\alpha := \hat{\alpha}$.

1. We check whether the inequality

$$f(x^k + \alpha d^k) \leq f(x^k) + \sigma\alpha \langle \nabla f(x^k), d^k \rangle \tag{2.4}$$

   whether it satisfies or not.

2. If (2.4) is not satisfied, we take $\alpha := \theta\alpha$ and return to Step 1.

   Otherwise, we accept $\alpha_k = \alpha$ as the step size value.

By constructing Armijo rule, α_k is the largest among all numbers of the form $\hat{\alpha}\theta^i$, $i = 0, 1, 2, ...$, which satisfies the inequality (2.4).

Since f is differentiable at point x^k, we have

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &\approx f(x^k) + \alpha \langle \nabla f(x^k), d^k \rangle - f(x^k) \\ &= \alpha \langle \nabla f(x^k), d^k \rangle, \end{aligned}$$

which gives us an interpretation for the number $\alpha\langle\nabla f(x^k), d^k\rangle$. It represents the estimate of real decrease, given by the linear approximation of f at the point x^k, for the step size α in the direction d^k. Therefore $\sigma\alpha\langle\nabla f(x^k), d^k\rangle$, in the inequality (2.4), is a fraction of this estimate (determined by $\sigma \in (0,1)$), then α_k is chosen so that the actual decrease in f is at least the fraction (determined by $\sigma \in (0,1)$) of the foreseen.

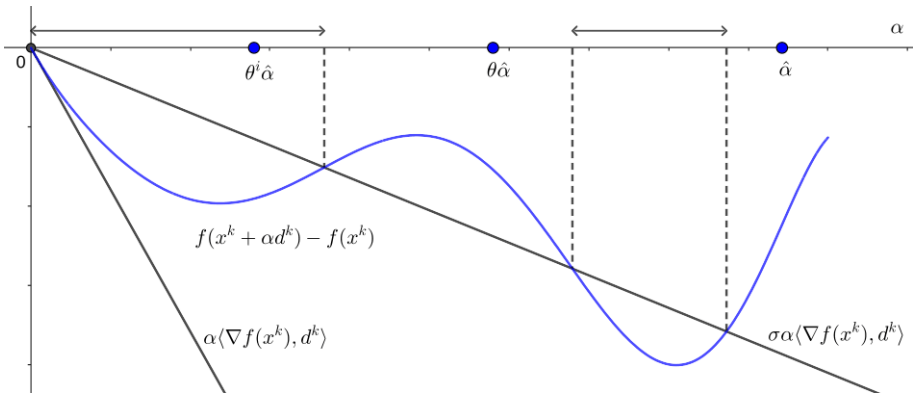


Figure 2.5: Illustration of the values of α that satisfy Armijo's rule and their respective images that satisfy the condition (2.4).

Step 1 of Armijo rule checks whether the parameter α provides a reasonable decrease in the function f, in order to satisfy the inequality (2.4). If the parameter α satisfies (2.4), then we accept it as the step size value. Otherwise, we use the θ parameter to reduce it and go back to step 1 to check if the new α provides the desired decrease. See the Figure 2.1.

The next lemma guarantees that Armijo rule is well defined and that this process ends in a finite number of iterations.

Lemma 2.1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function at the point $x^k \in \mathbb{R}^n$. Suppose that $d^k \in \mathbb{R}^n$ satisfies $\langle\nabla f(x^k), d^k\rangle < 0$. Then the inequality (2.4) is satisfied for all sufficiently small $\alpha > 0$. In particular, Armijo rule is well defined and ends with a $\alpha_k > 0$.*

Proof. For all $\alpha > 0$, as f is differentiable at the point $x^k \in \mathbb{R}^n$, we have that

$$f(x^k + \alpha d^k) = f(x^k) + \langle\nabla f(x^k), \alpha d^k\rangle + o(\alpha),$$

with $\lim_{\alpha\to 0_+} \frac{o(\alpha)}{\alpha} = 0$. Thus,

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &= \alpha\langle \nabla f(x^k), d^k\rangle + o(\alpha) \\ &= \sigma\alpha\langle \nabla f(x^k), d^k\rangle + (1-\sigma)\alpha\langle \nabla f(x^k), d^k\rangle + o(\alpha) \\ &= \sigma\alpha\langle \nabla f(x^k), d^k\rangle + \alpha\left((1-\sigma)\langle \nabla f(x^k), d^k\rangle + \frac{o(\alpha)}{\alpha}\right). \end{aligned}$$

Since $\lim_{\alpha\to 0_+} \frac{o(\alpha)}{\alpha} = 0$ and $0 < -\frac{(1-\sigma)}{2}\langle \nabla f(x^k), d^k\rangle$, there exists $\delta > 0$ such that if $\alpha \in (0, \delta)$,

$$\frac{o(\alpha)}{\alpha} \leq -\frac{(1-\sigma)}{2}\langle \nabla f(x^k), d^k\rangle.$$

Thus,

$$\begin{aligned} (1-\sigma)\langle \nabla f(x^k), d^k\rangle + \frac{o(\alpha)}{\alpha} &\leq (1-\sigma)\langle \nabla f(x^k), d^k\rangle - \frac{(1-\sigma)}{2}\langle \nabla f(x^k), d^k\rangle \\ &= \frac{(1-\sigma)}{2}\langle \nabla f(x^k), d^k\rangle \\ &< 0, \quad \forall\ \alpha \in (0, \delta). \end{aligned}$$

Therefore,

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &= \sigma\alpha\langle \nabla f(x^k), d^k\rangle + \alpha\left((1-\sigma)\langle \nabla f(x^k), d^k\rangle + \frac{o(\alpha)}{\alpha}\right) \\ &\leq \sigma\alpha\langle \nabla f(x^k), d^k\rangle, \quad \forall \alpha \in (0, \delta). \end{aligned}$$

This shows that the inequality (2.4) is satisfied for all sufficiently small α.
As $\hat{\alpha}\theta^i \to 0$, when $i \to \infty$, there exists $i_0 \in \mathbb{N} \cup \{0\}$ such that if $i \geq i_0$ then $\hat{\alpha}\theta^i \in (0, \delta)$, that is, all these $\hat{\alpha}\theta^i$ satisfy inequality (2.4) with the largest of them being $\hat{\alpha}\theta^{i_0}$. Now it remains to choose α_k as the largest of the numbers $\hat{\alpha}\theta^i$, $i = 0, 1, ..., i_0$, which satisfies the inequality (2.4). This shows that, in particular, Armijo rule is well defined and ends with a $\alpha_k > 0$. □

An advantage of using Armijo rule is that, in each iteration of the method, we choose the step size α_k without having to solve a one-dimensional minimization subproblem and with the guarantee that α_k provides a reasonable decrease of the f function. A disadvantage is that, for each k, as we need to check whether the inequality (2.4) is satisfied or not, it is necessary to evaluate the function f at the corresponding points and calculate $\sigma\alpha\langle \nabla f(x^k), d^k\rangle$, if evaluating the function f has a high computational cost, then this can considerably increase the time to obtain the step α_k .

The Lemma 2.1.2 guarantees that for every sufficiently small α, Armijo rule is well defined. If we add the hypothesis that the gradient of the function f is Lipschitz, then

for every k, it is possible to obtain a constant $\bar{\alpha}_k$ that guarantees that every $\alpha \in (0, \bar{\alpha}_k]$ satisfies the inequality (2.4).

Lemma 2.1.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on $\mathbb{R}^n$, with Lipschitz-continuous gradient on $\mathbb{R}^n$ with constant $L > 0$.*

If $x^k, d^k \in \mathbb{R}^n$ satisfy $\langle \nabla f(x^k), d^k \rangle < 0$, then the inequality (2.4) is valid for all $\alpha \in (0, \bar{\alpha}_k]$, where

$$\bar{\alpha}_k = \frac{2(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{L\|d^k\|^2} > 0. \tag{2.5}$$

Proof. By the Proposition 1.0.2, for all $\alpha \in \mathbb{R}$, we have that

$$\left| f(x^k + \alpha d^k) - f(x^k) - \langle \nabla f(x^k), \alpha d^k \rangle \right| \leq \frac{L}{2}\|\alpha d^k\|^2.$$

Thus,

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &\leq \alpha \langle \nabla f(x^k), d^k \rangle + \frac{L}{2}\alpha^2 \|d^k\|^2 \\ &= \alpha \left(\langle \nabla f(x^k), d^k \rangle + \frac{L}{2}\alpha \|d^k\|^2 \right). \end{aligned}$$

Therefore, for all $\alpha \in (0, \bar{\alpha}_k]$,

$$\begin{aligned} f(x^k + \alpha d^k) - f(x^k) &\leq \alpha \left(\langle \nabla f(x^k), d^k \rangle + \frac{L}{2}\bar{\alpha}_k \|d^k\|^2 \right) \\ &= \alpha \left(\langle \nabla f(x^k), d^k \rangle + \frac{L}{2} \frac{2(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{L\|d^k\|^2} \|d^k\|^2 \right) \\ &= \alpha \left(\langle \nabla f(x^k), d^k \rangle + (\sigma - 1)\langle \nabla f(x^k), d^k \rangle \right) \\ &= \sigma \alpha \langle \nabla f(x^k), d^k \rangle. \end{aligned}$$

□

We use the Proposition 1.0.2 in the Lemma 2.1.3 to estimate $f(x^k + \alpha d^k) - f(x^k)$ by $\alpha \langle \nabla f(x^k), d^k \rangle + \frac{L}{2}\alpha^2 \|d^k\|^2$. Note that if $\alpha \langle \nabla f(x^k), d^k \rangle + \frac{L}{2}\alpha^2 \|d^k\|^2 \leq \sigma \alpha \langle \nabla f(x^k), d^k \rangle$, then

$$\frac{L\alpha\|d^k\|^2}{2} \leq (\sigma - 1)\langle \nabla f(x^k), d^k \rangle,$$

thus,

$$\alpha \leq \frac{2(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{L\|d^k\|^2},$$

which gives an idea for choosing $\bar{\alpha}_k$.

Under the hypotheses of the Lemma 2.1.3, if

$$\frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|^2} \leq \delta < 0, \tag{2.6}$$

where δ is a constant that does not depend on k and if the parameters $\hat{\alpha}$, σ and θ are the same for each iteration, multiplying the inequality (2.6) by $(\sigma - 1)$, we have that

$$0 < \delta(\sigma - 1) \leq \frac{(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{\|d^k\|^2},$$

now multiplying the inequality by $\frac{2}{L}$,

$$0 < \frac{2\delta(\sigma - 1)}{L} \leq \frac{2(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{L\|d^k\|^2},$$

that is,

$$0 < \frac{-2\delta(1 - \sigma)}{L} \leq \frac{2(\sigma - 1)\langle \nabla f(x^k), d^k \rangle}{L\|d^k\|^2},$$

which means,

$$0 < \bar{\alpha} \leq \bar{\alpha_k}, \quad \forall k, \quad \text{where} \quad \bar{\alpha} := \frac{-2\delta(1 - \sigma)}{L} > 0.$$

Therefore, the inequality (2.4) is satisfied for all $\alpha \in (0, \bar{\alpha}]$. We know that α_k is the largest number of the form $\hat{\alpha}\theta^i$, $i = 0, 1, 2, ...$, which satisfies the inequality (2.4). Since step size value greater than α_k was not accepted, then either $i = 0$ (and in this case α_k is the largest allowed value, i.e. $\alpha_k = \hat{\alpha}$) or $i \neq 0$ (and in this case $\frac{\alpha_k}{\theta} = \frac{\hat{\alpha}\theta^i}{\theta} = \hat{\alpha}\theta^{i-1}$ was not accepted), thus

$$\text{either} \quad \alpha_k = \hat{\alpha} \quad \text{or} \quad \frac{\alpha_k}{\theta} > \bar{\alpha},$$

that is,

$$\text{either} \quad \alpha_k = \hat{\alpha} \quad \text{or} \quad \alpha_k > \theta\bar{\alpha}.$$

therefore,

$$\alpha_k \geq \min\{\hat{\alpha}, \theta\bar{\alpha}\} := \check{\alpha} > 0, \quad \forall k.$$

Fixed step size rule. We fix a number $\hat{t} > 0$ that does not depend on k and take $t_k = \hat{t}$ for all iterations.

An advantage of using the fixed step size rule is that it is the simplest rule among those presented. On the other hand, this causes the rule to have major drawbacks, as prefixing a step size value for all iterations causes "best" step lengths to be ignored. Furthermore, if the size of fixed step is too large, this may result in the method not converging; and if it is too small, then convergence may be very slow (considerably increasing the number of method iterations).

Remark 2.1.2. *Assuming that the hypotheses of the Lemma 2.1.3 and the condition (2.6) hold, then Armijo inequality (2.4) will hold for $\hat{t}$, if $\hat{t} \in (0, \bar{\alpha}]$. In Armijo rule, defining*

$\hat{\alpha} := \hat{t}$, we obtain that $\alpha_1 = \hat{t}$, $\alpha_2 = \hat{t}$, $\alpha_3 = \hat{t}$, and so on. This shows that the convergence of methods with sufficiently fixed step size small follows from the convergence of methods using Armijo rule.

2.2 Gradient method

The three most important properties of the gradient of a differentiable function are as follows: Given $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$, then

1. The gradient is a direction in which the function f is increasing.
2. Among all the directions along which the function f grows, the direction of the gradient is the fastest growing.
3. The gradient of f at point x is perpendicular to the level surface of f that passes through that point.

The above results are well known and can be found, for example, in [9] and [14].

This motivates the definition of the gradient method to solve (2.1), since $-\nabla f(x)$ is a direction of descent of the function f at the point x and is the direction of decrease most fast.

Suppose that the function f is differentiable in $\mathbb{R}^n$. In the context of descent methods, the gradient method is, by definition,

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \ k = 0, 1, 2, ..., \tag{2.7}$$

that is, we take the direction of descent $d^k = -\nabla f(x^k)$, for all k. If $\nabla f(x^k) = 0$, for some k, x^k is a critical point of the Problem (2.1) and the method stops.

Algorithm 2 Gradient method

1: Choose $x^0 \in \mathbb{R}^n$ and set $k := 0$;
2: **while** $||\nabla f(x^k)|| \neq 0$ **do**
3: Choose $d^k = -\nabla f(x^k)$;
4: Choose $t_k > 0$ such that $f(x^k + t_k d^k) < f(x^k)$, using one of the three search rules presented (one-dimensional minimization rule, Armijo rule or fixed step size rule);
5: Set $x^{k+1} := x^k + t_k d^k$;
6: Update $k := k + 1$;
7: **end while**

The gradient method using one-dimensional minimization is called the maximum descent method. Since we take $d^k = -\nabla f(x^k)$ and use one-dimensional minimization, by (2.3), it follows that

$$\langle \nabla f(x^{k+1}), \nabla f(x^k) \rangle = 0, \tag{2.8}$$

therefore, the directions used in subsequent iterations are orthogonal. Then, by (2.7) and (2.8), see that for arbitrary k,

$$\begin{aligned} \langle x^{k+2} - x^{k+1}, x^{k+1} - x^k \rangle &= \langle -\alpha_{k+1} \nabla f(x^{k+1}), -\alpha_k \nabla f(x^k) \rangle \\ &= \alpha_{k+1} \alpha_k \langle \nabla f(x^{k+1}), \nabla f(x^k) \rangle \\ &= 0, \end{aligned}$$

therefore,

$$(x^{k+2} - x^{k+1}) \perp (x^{k+1} - x^k). \tag{2.9}$$

This justifies and illustrates the fact that the method has a "zig-zag" trajectory throughout the iterations. See the figure 2.2 below.

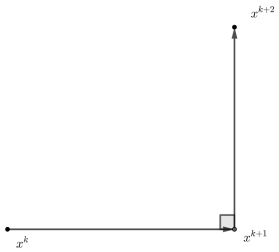


Figure 2.6: Illustration of property (2.9).

Armijo inequality (2.4) is given by

$$f(x^k - \alpha_k \nabla f(x^k)) \leq f(x^k) - \sigma \alpha_k \|\nabla f(x^k)\|^2. \tag{2.10}$$

When $\nabla f(x^k) \neq 0$, we have

$$\frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|^2} = \frac{\langle \nabla f(x^k), -\nabla f(x^k) \rangle}{\|-\nabla f(x^k)\|^2} = -1 < 0,$$

in this case, the condition (2.6) is satisfied with $\delta = -1$ and the estimate (2.5) of longest step is given by

$$\bar{\alpha_k} = \frac{2(1-\sigma)}{L} > 0. \tag{2.11}$$

As $\bar{\alpha_k}$ does not depend on k, when the gradient of the function f is Lipschitz continuous in $\mathbb{R}^n$ and the Armijo rule is being used, at least Lemma 2.1.3 and subsequent comments, we have

$$\alpha_k \geq \check{\alpha} > 0, \tag{2.12}$$

where $\check{\alpha}$ does not depend on k.

Theorem 2.2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on $\mathbb{R}^n$, with Lipschitz-continuous gradient on $\mathbb{R}^n$ with module $L > 0$. In the case where the Algorithm 2 uses a fixed step size, let us assume that $\hat{t}$ is such that*

$$0 < \hat{t} < \frac{2}{L}. \tag{2.13}$$

Then, if a sequence $\{x^k\}$ generated by the Algorithm 2 has a cluster point, or if the function f is lower bound in $\mathbb{R}^n$, we have that

$$\{\nabla f(x^k)\} \to 0 \quad (k \to \infty). \tag{2.14}$$

In particular, each cluster point of any sequence $\{x^k\}$ generated by the Algorithm 2 is a critical point of the Problem (2.1).

Proof. Let us first consider the case of Armijo rule. If $\nabla f(x^k) \neq 0$ for all k, then the sequence $\{f(x^k)\}$ is decreasing. Suppose that the sequence $\{x^k\}$ has a cluster point. Then there is a subsequence $\{x^{k_j}\} \to \bar{x}$, when $j \to \infty$. Due to the continuity of f, we have $\lim_{j\to\infty} f(x^{k_j}) = f(\bar{x})$, therefore $f(\bar{x})$ is a cluster point of the sequence $\{f(x^k)\}$. Since $\{f(x^k)\}$ is monotone and has a bounded subsequence $\{f(x^{k_j})\}$, then $\{f(x^k)\}$ is bounded too. Thus, as $\{f(x^k)\}$ is monotone and bounded, $\{f(x^k)\}$ is convergent. If the function f is lower bounded on $\mathbb{R}^n$, then $\{f(x^k)\}$ is convergent (even if $\{x^k\}$ does not have cluster points). By the Armijo inequality (2.10) and (2.12), for all k, we have

$$f(x^k) - f(x^{k+1}) \geq \sigma\alpha_k\|\nabla f(x^k)\|^2 \geq \sigma\check{\alpha}\|\nabla f(x^k)\|^2 > 0. \tag{2.15}$$

Since $\lim_{k\to\infty} f(x^k) - f(x^{k+1}) = 0$, we have $\lim_{k\to\infty} \sigma\check{\alpha}\|\nabla f(x^k)\|^2 = 0$. Therefore,

$$\lim_{k\to\infty} \nabla f(x^k) = 0$$

which guarantees (2.14). In particular, if $\{x^{k_j}\}$ converges to $\bar{x}$, as the gradient of the function f is continuous on $\mathbb{R}^n$, it follows that

$$0 = \lim_{j\to\infty} \nabla f(x^{k_j}) = \nabla f(\bar{x}),$$

that is, $\bar{x}$ is a critical point of the Problem (2.1).

Let us now consider the case of the one-dimensional minimization rule. For all k, let us denote by $\tilde{x}^{k+1}$ the point that would be obtained by Armijo rule, with $\tilde{\alpha}_k$ being the associated step size. By the definition of x^{k+1}, we have

$$f(x^k) - f(x^{k+1}) \geq f(x^k) - f(\tilde{x}^{k+1}) \geq \sigma\tilde{\alpha}_k\|\nabla f(x^k)\|^2 \geq \sigma\check{\alpha}\|\nabla f(x^k)\|^2 > 0.$$

Thus, the result follows from the previous analysis by replacing α_k with $\tilde{\alpha}_k$.

Now considering the case of the fixed step size rule, we take $\alpha_k = \hat{t}$ for all iterations. By hypothesis, (2.13), we are assuming that $\hat{t} < \frac{2}{L}$. In the gradient method, we saw that the longest step estimate is given by the condition (2.11), as we take $\alpha_k = \hat{t} < \frac{2}{L}$, for σ sufficiently small, we have

$$\alpha_k = \hat{t} < \frac{2}{L}(1-\sigma) < \frac{2}{L}, \quad \forall k,$$

that is, the choice of step size belongs to the set $(0, \bar{\alpha_k}]$, $\forall k$. This shows that if $\hat{t} < \frac{2}{L}$ and we do $\hat{\alpha} := \hat{t}$ in Armijo rule, with σ small enough, using the fixed step size rule is equivalent to using Armijo rule where in each iteration the inequality (2.4) is already true for $\alpha = \hat{t}$. Therefore, the result follows from the analysis made in the case of the Armijo Rule. □

In the proof of the Theorem 2.2.1, particularly in the inequality (2.15), the importance of the step size α_k not tending to 0 becomes evident. It is possible to exchange the hypothesis that the gradient of f is Lipschitz-continuous with the weaker hypothesis that it is continuous. In this case, the condition (2.12) may not happen and this makes the argument made in (2.15) impossible.

Theorem 2.2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function in $\mathbb{R}^n$, with a continuous gradient. Let's suppose that the Algorithm 2 uses one-dimensional minimization or Armijo rule.*

Then each cluster point of any sequence $\{x^k\}$ generated by the Algorithm 2 is a critical point of the Problem (2.1).

Proof. The case in which the Algorithm 2 uses the one-dimensional minimization rule reduces to the case of Armijo rule in the same way as in the demonstration of the Theorem 2.2.1. Therefore, let us consider the Armijo rule.

Suppose that the Algorithm 2 uses the Armijo rule and $\{x^k\}$ is a generated sequence. Suppose that $\{x^k\}$ has an cluster point $\bar{x} \in \mathbb{R}^n$ and that $\nabla f(x^k) \neq 0$ for all k. Then there is a subsequence $\{x^{k_j}\}$ that converges to $\bar{x}$ and $\{f(x^k)\}$ is decreasing.

Suppose that there exists $\check{\alpha} > 0$ such that $\alpha_{k_j} \geq \check{\alpha}$, for all j. Due to the continuity of f, we have that $\lim_{j\to\infty} f(x^{k_j}) = f(\bar{x})$, hence $f(\bar{x})$ is a cluster point of the sequence $\{f(x^k)\}$. Since $\{f(x^k)\}$ is monotone and has a bounded subsequence $\{f(x^{k_j})\}$, it follows that $\{f(x^k)\}$ is bounded. Thus, as $\{f(x^k)\}$ is monotone and bounded, $\{f(x^k)\}$ is convergent. Using that $\{f(x^k)\}$ is decreasing and Armijo inequality, for all j, we obtain that

$$\begin{aligned} f(x^{k_{j+1}}) &\leq f(x^{k_{j+1}-1}) \leq \ldots \leq f(x^{k_j+1}) \\ &\leq f(x^{k_j}) - \sigma\alpha_{k_j}\|\nabla f(x^{k_j})\|^2, \end{aligned}$$

thus,

$$f(x^{k_j}) - f(x^{k_{j+1}}) \geq \sigma\alpha_{k_j}\|\nabla f(x^{k_j})\|^2 \geq \sigma\check{\alpha}\|\nabla f(x^{k_j})\|^2 > 0.$$

Since $\lim_{j\to\infty} f(x^{k_j}) - f(x^{k_{j+1}}) = 0$, we have $\lim_{j\to\infty} \sigma\check{\alpha}\|\nabla f(x^{k_j})\|^2 = 0$. Therefore,

$$\lim_{j\to\infty} \nabla f(x^{k_j}) = 0.$$

Since the gradient of the function f is continuous in $\mathbb{R}^n$ and the subsequence $\{x^{k_j}\}$ converges to $\bar{x}$, we have

$$0 = \lim_{j\to\infty} \nabla f(x^{k_j}) = \nabla f(\bar{x}),$$

therefore $\bar{x}$ is a critical point of the Problem (2.1).

Suppose that there is no $\check{\alpha} > 0$ such that $\alpha_{k_j} \geq \check{\alpha}$, for all j. Then for every $\check{\alpha} > 0$ there is a $j \in \mathbb{N}$ such that $\alpha_{k_j} < \check{\alpha}$. In particular, for every $i \in \mathbb{N}$, such as $\frac{1}{i} > 0$, there exists $j_i \in \mathbb{N}$ such that $\alpha_{k_{j_i}} < \frac{1}{i}$. Thus, the sequence $\{\alpha_{k_j}\}_{j\in\mathbb{N}}$ has a subsequence $\{\alpha_{k_{j_i}}\}_{i\in\mathbb{N}}$ which converges to 0, when $i \to \infty$. Therefore, without loss of generality, we will assume that $\{\alpha_{k_j}\} \to 0$, when $j \to \infty$. Combining this information and step 2 of Armijo rule, for every sufficiently large j, the initial value of the step size $\hat{\alpha}$ was reduced at least once. Therefore, $\frac{\alpha_{k_j}}{\theta}$ does not satisfy Armijo inequality (2.4), that is,

$$f\left(x^{k_j} - \frac{\alpha_{k_j}}{\theta}\nabla f(x^{k_j})\right) > f(x^{k_j}) - \sigma\frac{\alpha_{k_j}}{\theta}\|\nabla f(x^{k_j})\|^2.$$

denoting $\tilde{\alpha}_{k_j} = \frac{\alpha_{k_j}}{\theta}\|\nabla f(x^{k_j})\|$ and dividing the last inequality by $\tilde{\alpha}_{k_j}$, we have

$$\frac{f\left(x^{k_j} - \frac{\alpha_{k_j}}{\theta}\nabla f(x^{k_j})\right) - f(x^{k_j})}{\tilde{\alpha}_{k_j}} > -\sigma\|\nabla f(x^{k_j})\|,$$

or even,

$$\frac{f\left(x^{k_j} - \tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right) - f(x^{k_j})}{\tilde{\alpha}_{k_j}} > -\sigma\|\nabla f(x^{k_j})\|. \tag{2.16}$$

Applying the Mean Value Theorem 1.0.3, for each j there exists $t_{k_j} \in [0,1]$ such that

$$\begin{aligned}
&f\left(x^{k_j} - \tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right) - f(x^{k_j}) = \\
&= \left\langle \nabla f\left(t_{k_j}x^{k_j} + (1-t_{k_j})(x^{k_j} - \tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|})\right), x^{k_j} - \tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|} - x^{k_j}\right\rangle \\
&= \left\langle \nabla f\left(x^{k_j} - \tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|} + t_{k_j}\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right), -\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right\rangle \\
&= \left\langle \nabla f\left(x^{k_j} - (1-t_{k_j})\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right), -\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right\rangle.
\end{aligned} \tag{2.17}$$

Combining equality (2.17) and inequality (2.16), we have

$$\frac{\left\langle \nabla f\left(x^{k_j} - (1-t_{k_j})\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right), -\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right\rangle}{\tilde{\alpha}_{k_j}} > -\sigma\|\nabla f(x^{k_j})\|,$$

that is,

$$-\left\langle \nabla f\left(x^{k_j} - (1-t_{k_j})\tilde{\alpha}_{k_j}\frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right), \frac{\nabla f(x^{k_j})}{\|\nabla f(x^{k_j})\|}\right\rangle > -\sigma\|\nabla f(x^{k_j})\|. \tag{2.18}$$

Since $\{\alpha_{k_j}\} \to 0$, $\{t_{k_j}\}$ is bounded and $\{\frac{\|\nabla f(x^{k_j})\|}{\theta}\}$ is bounded (because $\nabla f(x^{k_j}) \to \nabla f(\bar{x})$), then $\tilde{\alpha}_{k_j} = \frac{\alpha_{k_j}}{\theta}\|\nabla f(x^{k_j})\| \to 0$ $(j \to \infty)$. Turning to the limit when $j \to \infty$ in the relation (2.18), we have

$$-\left\langle \nabla f(\bar{x}), \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}\right\rangle \geq -\sigma\|\nabla f(\bar{x})\|$$

or,

$$-\|\nabla f(\bar{x})\| \geq -\sigma\|\nabla f(\bar{x})\|.$$

Since $\sigma \in (0,1)$, then $\nabla f(\bar{x}) = 0$, so $\bar{x}$ is a critical point of the Problem 2.1. □

Remark 2.2.1. *A natural hypothesis is to assume that the level set $L_{f,\mathbb{R}^n}(f(x^0))$ is bounded. If it is bounded, as we are working with descent methods, $f(x^k) \leq f(x^0)$, for all k, therefore the sequence $\{x^k\} \subset L(f(x^0))$ and will also be bounded. This guarantees that $\{x^k\}$ admits at least one cluster point and, according to the Theorems 2.2.1 and 2.2.2, it will be a critical point.*

In Theorems 2.2.1 and 2.2.2, we proved that if the sequence $\{x^k\}$ has cluster points, they are critical points of the Problem (2.1). If we add the assumption that the function f is convex, then we obtain stronger convergence results. In this case, if the set of minimizers is non-empty, then the sequence $\{x^k\}$ converges to a solution of the Problem (2.1).

Theorem 2.2.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, differentiable in $\mathbb{R}^n$, with continuous gradient. Suppose that the Algorithm 2 uses the Armijo rule with $\hat{\alpha} \leq 1$. If the set of unconstrained minimizers of f is non-empty, then any sequence $\{x^k\}$ generated by the Algorithm 2 converges to a solution of the Problem (2.1).*

Proof. Let $\bar{x} \in \mathbb{R}^n$ be a solution to the problem. By Armijo inequality (2.10), for all k, we have

$$f(x^k) - f(x^{k+1}) \geq \sigma \alpha_k \|\nabla f(x^k)\|^2.$$

Then,

$$\begin{aligned} f(x^0) - f(\bar{x}) &\geq f(x^0) - f(x^k) \\ &= f(x^0) - f(x^1) + f(x^1) - f(x^2) + f(x^2) - \dots \\ &\quad - f(x^{k-2}) + f(x^{k-2}) - f(x^{k-1}) + f(x^{k-1}) - f(x^k) \\ &= \sum_{i=0}^{k-1} (f(x^i) - f(x^{i+1})) \\ &\geq \sigma \sum_{i=0}^{k-1} \alpha_i \|\nabla f(x^i)\|^2. \end{aligned}$$

Passing the limit when $k \to \infty$, we have

$$\sum_{i=0}^{\infty} \alpha_i \|\nabla f(x^i)\|^2 \leq \frac{f(x^0) - f(\bar{x})}{\sigma}. \tag{2.19}$$

Due to the convexity of the function f, the Proposition 1.0.4 and the optimality of $\bar{x}$, we have that

$$\langle \nabla f(x^k), \bar{x} - x^k \rangle \leq f(\bar{x}) - f(x^k) \leq 0. \tag{2.20}$$

Using (2.7), (2.20) and the relation $\alpha_k^2 \leq \alpha_k$, we have that

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|^2 &= \|x^k - \bar{x} + x^{k+1} - x^k\|^2 \\
&= \langle x^k - \bar{x} + x^{k+1} - x^k, x^k - \bar{x} + x^{k+1} - x^k \rangle \\
&= \|x^k - \bar{x}\|^2 + 2\langle x^k - \bar{x}, x^{k+1} - x^k \rangle + \|x^{k+1} - x^k\|^2 \\
&= \|x^k - \bar{x}\|^2 + 2\langle x^k - \bar{x}, -\alpha_k \nabla f(x^k) \rangle + \|-\alpha_k \nabla f(x^k)\|^2 \\
&= \|x^k - \bar{x}\|^2 + 2\alpha_k \langle \nabla f(x^k), \bar{x} - x^k \rangle + \alpha_k^2 \|\nabla f(x^k)\|^2 \\
&\leq \|x^k - \bar{x}\|^2 + \alpha_k \|\nabla f(x^k)\|^2. \qquad (2.21)
\end{aligned}$$

Let us fix an arbitrary k. Using the inequality (2.21) in chain, for all $j \geq k+1$, we have that

$$\begin{aligned}
\|x^j - \bar{x}\|^2 &\leq \|x^{j-1} - \bar{x}\|^2 + \alpha_{j-1} \|\nabla f(x^{j-1})\|^2 \\
&\leq \|x^{j-2} - \bar{x}\|^2 + \alpha_{j-2} \|\nabla f(x^{j-2})\|^2 + \alpha_{j-1} \|\nabla f(x^{j-1})\|^2 \\
&\vdots \\
&\leq \|x^k - \bar{x}\|^2 + \sum_{i=k}^{j-1} \alpha_i \|\nabla f(x^i)\|^2 \qquad (2.22) \\
&\leq \|x^k - \bar{x}\|^2 + \sum_{i=0}^{\infty} \alpha_i \|\nabla f(x^i)\|^2 < +\infty,
\end{aligned}$$

where we used (2.19) in the last inequality. This shows that the sequence $\{x^k\}$ is bounded, so $\{x^k\}$ has an cluster point $\hat{x}$. Thus, Theorem 2.2.2 guarantees that $\nabla f(\hat{x}) = 0$. Using convexity, we conclude that $\hat{x}$ is a solution to the Problem (2.1) (by Theorem 2.2.2). So the same analysis for $\bar{x}$ can be done for $\hat{x}$. Therefore, from (2.22), for all $j \geq k+1$, we have

$$\|x^j - \hat{x}\|^2 \leq \|x^k - \hat{x}\|^2 + \sum_{i=k}^{\infty} \alpha_i \|\nabla f(x^i)\|^2, \qquad (2.23)$$

and from (2.19), we have

$$\lim_{k \to \infty} \left(\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x^i)\|^2 \right) = 0.$$

Given $\delta > 0$ arbitrarily small, there exists $k_1 \in \mathbb{N}$ such that if $k > k_1$, then

$$\sum_{i=k}^{\infty} \alpha_i \|\nabla f(x^i)\|^2 < \frac{\delta}{2}.$$

Since $\hat{x}$ is an cluster point of the sequence $\{x^k\}$, there exists $k_2 \in \mathbb{N}$ such that $k_2 > k_1$ and

$$\|x^{k_2} - \hat{x}\|^2 < \frac{\delta}{2}.$$

From the relation (2.23), for all $\delta > 0$,

$$\begin{aligned} \|x^j - \hat{x}\|^2 &\leq \|x^{k_2} - \hat{x}\|^2 + \sum_{i=k_2}^{\infty} \alpha_i \|\nabla f(x^i)\|^2 \\ &< \frac{\delta}{2} + \frac{\delta}{2} = \delta, \quad j \geq k_2 + 1. \end{aligned}$$

This proves that $\{x^k\}$ converges to $\hat{x}$. □

Chapter 3

Subgradient method

The chapter will be divided into two sections. In the first section, we will present the first changes that arise when we move from the differentiable context to the non-differentiable context. The main results are the fact that given $y \in \partial f(x)$, it can happen that $-y \notin \mathcal{D}_f(x)$, that in general we only know one subgradient at each point and the difficulty in choosing stopping rules. In the second section, we will introduce the subgradient method, which is not necessarily a descent method, in each iteration we take the next step in the opposite direction of a subgradient, and in the simplest cases the step size is pre-fixed and is not chosen using a line search. We will also present step size rules and perform a convergence analysis of the method. The main results are some inequalities that help in the proof of convergence and the proofs of convergence.

The construction of the first and second sections were based on references [2, 4, 16, 19, 21].

3.1 Non-differentiable convex optimization

Let us consider the unconstrained minimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n \end{aligned} \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function in $\mathbb{R}^n$. Therefore, f may not be differentiable.

By the Proposition 1.0.5, the directional derivative of the function f at the point $x \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$ satisfies the following condition

$$f'(x; d) = \max\{\langle y, d\rangle \mid y \in \partial f(x)\}, \tag{3.2}$$

where $\partial f(x)$ is the subdifferential of f at the point x defined in 1.0.5, that is,

$$\partial f(x) = \{y \in \mathbb{R}^n \mid f(z) \geq f(x) + \langle y, z - x\rangle, \quad \forall z \in \mathbb{R}^n\}. \tag{3.3}$$

Given $d \in \mathcal{D}_f(x)$, there exists $\varepsilon > 0$ such that $f(x + td) < f(x)$ for all $t \in (0, \varepsilon]$. Thus $\frac{f(x+td)-f(x)}{t} < 0$, for all $t \in (0, \varepsilon]$, therefore $f'(x; d) \leq 0$. Using the relation (3.2), we conclude that $\max\{\langle y, d\rangle \mid y \in \partial f(x)\} \leq 0$ and, therefore, that $\langle y, d\rangle \leq 0$ for all $y \in \partial f(x)$.

On the other hand, if $\langle y, d\rangle < 0$ for all $y \in \partial f(x)$ then $\max\{\langle y, d\rangle \mid y \in \partial f(x)\} < 0$. Using the relation (3.2), we conclude that $f'(x; d) < 0$ and therefore, there exists $\delta > 0$ such that $\frac{f(x+td)-f(x)}{t} < 0$ for all $t \in (0, \delta]$. Thus, $f(x + td) < f(x)$ for all $t \in (0, \delta]$ and therefore $d \in \mathcal{D}_f(x)$.

Thus, for $d \in \mathcal{D}_f(x)$ it is necessary that $\langle y, d\rangle \leq 0$ for all $y \in \partial f(x)$, and for $\langle y, d\rangle < 0$ for all $y \in \partial f(x)$ it is necessary that $d \in \mathcal{D}_f(x)$. In both cases, it is necessary to know the entire set $\partial f(x)$.

Next, we will see that given $y \in \partial f(x)$, it can happen that $-y \notin \mathcal{D}_f(x)$.

Example 3.1.1. *Let $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = |x_1| + 2|x_2|$. Let us consider $x = (x_1, 0)$, where $x_1 > 0$ is arbitrary. First, we will show that*

$$y = (1, 2) \in \partial f(x).$$

In fact, for all $z = (z_1, z_2) \in \mathbb{R}^2$, we have that

$$|z_1| + 2|z_2| \geq \ z_1 + 2z_2,$$

using $x = (x_1, 0)$, we have

$$|z_1| + 2|z_2| \geq \ |x_1| + z_1 - x_1 + 2z_2,$$

therefore, by defining the function f and manipulating the terms $z_1 - x_1 + 2z_2$, we can write

$$f((z_1, z_2)) \geq f((x_1, 0)) + \langle (1, 2), (z_1, z_2) - (x_1, 0)\rangle,$$

i.e,

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

Now, let's show that $-y \notin \mathcal{D}_f(x)$. *In fact, for every sufficiently small* $t > 0$,

$$\begin{aligned} f(x - ty) = f((x_1 - ty_1, x_2 - ty_2)) &= |x_1 - ty_1| + 2|x_2 - ty_2| = |x_1 - t| + 2|0 - 2t| \\ &= x_1 + 3t \\ &> x_1 = |x_1| + 2|x_2| = f(x). \end{aligned}$$

In general, we are able to evaluate the objective function at current points, we know a subgradient at each point and combinations of this information throughout the iterations. This means that it is not expected to know the entire set $\partial f(x)$ in each x and with this comes another difficulty in applying the ideas of descent methods in the non-differentiable context.

In the differentiable context, a common stopping criterion is given by the condition $\|\nabla f(x^k)\| \leq \varepsilon$, for some small tolerance $\varepsilon > 0$. This condition does not directly apply to the non-differentiable case because the set-point operator $x \to \partial f(x)$ is not "continuous", as we will see in the next example.

Example 3.1.2. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be given by* $f(x) = |x|$. *The point* $\overline{x} = 0$ *is the only unconstrained global minimizer of* f. *We have*

$$\partial f(x) = \begin{cases} -1, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ 1, & \text{if } x > 0. \end{cases}$$

Then for every point $x^k \neq 0$, *with* x^k *converging to* $\overline{x} = 0$, *we have that* $|y| = 1$ *for every* $y \in \partial f(x^k)$. *Therefore, even if* x^k *is close to the solution of the problem,* $\partial f(x^k)$ *does not have subgradients with a small norm. Furthermore, it may happen that we find the solution* $x^k = 0$ *for some* k, *and this fact is not recognized by the method if we only know one subgradient at each point, since we compute* $y \in \partial f(0) = [-1, 1]$ *and therefore* $|y|$ *can be nonzero.*

3.2 Subgradient method

Let us consider the following algorithm to solve the Problem (3.1), that is, to minimize a convex function.

Algorithm 3 Subgradient method

1: Choose a sequence $\{\alpha_k\} \subset \mathbb{R}_+$;

2: Choose $x^1 \in \mathbb{R}^n$ and set $k := 1$;

3: Compute $d^k \in \partial f(x^k)$;

4: Compute

$$x^{k+1} = x^k - \alpha_k d^k; \tag{3.4}$$

5: Set $k := k + 1$ and return to Step 3.

In step 1, we pre-fix a sequence of step sizes that will be taken in each iteration of the method. In step 2, we take any starting point $x^1 \in \mathbb{R}^n$ and set $k := 1$. In step 3, we calculate some subgradient $d^k \in \partial f(x^k)$, and, by the Proposition 1.0.5, the set $\partial f(x^k)$ is non-empty. We are also assuming that it is possible to compute a subgradient at each point. In step 4, we use the point x^k, the step size α_k and the subgradient d^k to obtain the point x^{k+1}. In step 5, we do $k := k + 1$ and repeat the procedure for the new point obtained.

According to the structure of the method, from x^k, we calculate a subgradient d^k in the set $\partial f(x^k)$ and use the step size α_k to walk in the opposite direction to the subgradient. Therefore, the subgradient method looks like the gradient method for differentiable functions.

In the simplest cases, the sequence of step sizes $\{\alpha_k\}$ is pre-fixed at step 1 and the step size are not chosen using a line search.

As we saw in the example 3.1.1, given $y \in \partial f(x)$ it can happen that $-y \notin \mathcal{D}_f(x)$. Therefore, by building the Algorithm 3, the subgradient method is not necessarily a descent method.

It is common that throughout the iterations, we keep the "best" point obtained so far, that is, the point that provides the lowest value f so far. Thus, we define $f^1_{best} = f(x^1)$ and for $k \geq 2$,

$$f^k_{best} = \min\{f^{k-1}_{best}, f(x^k)\}.$$

Therefore,

$$f^k_{best} = \min\{f(x^1), \ldots, f(x^k)\},$$

that is, from a finite amount of points $x^1, \ldots, x^k$ obtained, we have a finite amount of images $f(x^1), \ldots, f(x^k)$, and we take f^k_{best} as the smallest of them all. As a consequence,

the sequence $\{f_{best}^k\}$ is non-increasing.

Step size rules. Let us cover 5 basic step length rules. Specifically, the rules of constant step size, constant step length, square summable but not summable, nonsummable diminishing and nonsummable diminishing step lengths.

Constant step size. We choose the step size $\alpha_k = \alpha$ for all iterations, where $\alpha > 0$ does not depend on k.

Constant step length. We choose the step size $\alpha_k = \frac{\gamma}{\|d^k\|}$, for all iterations, where $\gamma > 0$ does not depend on k.

Square summable but not summable. We choose the sequence $\{\alpha_k\}$ of step sizes such that it satisfies the following conditions:

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = +\infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < +\infty.$$

For example, $\alpha_k = 1/k$, for all k.

Nonsummable diminishing. We choose the sequence $\{\alpha_k\}$ of step size such that it satisfies the following conditions:

$$\alpha_k \geq 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = +\infty.$$

For example, $\alpha_k = 1/\sqrt{k}$, for all k.

Nonsummable diminishing step lengths. We choose the step size $\alpha_k = \frac{\gamma_k}{\|d^k\|}$ for all iterations, where the sequence $\{\gamma_k\}$ satisfies the following conditions:

$$\gamma_k \geq 0, \quad \lim_{k\to\infty} \gamma_k = 0, \quad \sum_{k=1}^{\infty} \gamma_k = +\infty.$$

When using one of these 5 step length rules, it is defined and pre-fixed in step 1 of the Algorithm 3, that is, it does not depend on data obtained during the iterations. Unlike line search, which depends on the current point and the fixed descent direction.

Convergence analysis. We will prove convergence results for each of the 5 step size rules presented. We will see that in the case of the constant step size and constant step length rules we guarantee that f_{best}^k converges to an interval "close" to the solution $f(\bar{x})$, and that in the case of the rules square summable but not summable, nonsummable diminishing and nonsummable diminishing step lengths we guarantee that f_{best}^k converges to the solution $f(\bar{x})$.

For the analysis, we assume that there is a minimizer of f, say $\bar{x}$. We assume that there is a L such that $\|d^k\| \leq L$, for all k; this condition holds, for example, when the function f is Lipschitz. We assume that we know $R > 0$ such that $R \geq \|x^1 - \bar{x}\|$.

Next, we will prove some classic inequalities that will be useful in convergence proofs.

By defining the point x^{k+1} in the relation 3.4, we obtain that

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|^2 &= \|x^k - \alpha_k d^k - \bar{x}\|^2 \\
&= \|x^k - \bar{x} - \alpha_k d^k\|^2 \\
&= \langle x^k - \bar{x} - \alpha_k d^k, x^k - \bar{x} - \alpha_k d^k \rangle \\
&= \|x^k - \bar{x}\|^2 - \alpha_k \langle x^k - \bar{x}, d^k \rangle - \alpha_k \langle d^k, x^k - \bar{x} \rangle + \|-\alpha_k d^k\|^2 \\
&= \|x^k - \bar{x}\|^2 - 2\alpha_k \langle d^k, x^k - \bar{x} \rangle + \alpha_k^2 \|d^k\|^2. \qquad (3.5)
\end{aligned}$$

Since $d^k \in \partial f(x^k)$, by the definition of the subdifferential (3.3), we have

$$f(\bar{x}) \geq f(x^k) + \langle d^k, \bar{x} - x^k \rangle,$$

thus,

$$f(\bar{x}) - f(x^k) \geq \langle d^k, \bar{x} - x^k \rangle,$$

or even,

$$f(x^k) - f(\bar{x}) \leq \langle d^k, x^k - \bar{x} \rangle. \qquad (3.6)$$

Combining the equality (3.5) with the inequality (3.6), we obtain that

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^k - \bar{x}\|^2 - 2\alpha_k \left(f(x^k) - f(\bar{x}) \right) + \alpha_k^2 \|d^k\|^2. \qquad (3.7)$$

This inequality (3.7) is the property that makes subgradient methods work. From there, for sufficiently small steps, the distance to the set of solutions decreases.

Applying the inequality (3.7) as a chain, we have

$$\begin{aligned}
\|x^{k+1} - \bar{x}\|^2 &\leq \|x^k - \bar{x}\|^2 - 2\alpha_k \left(f(x^k) - f(\bar{x}) \right) + \alpha_k^2 \|d^k\|^2 \\
&\leq \|x^{k-1} - \bar{x}\|^2 - 2\alpha_{k-1} \left(f(x^{k-1}) - f(\bar{x}) \right) + \alpha_{k-1}^2 \|d^{k-1}\|^2 \\
&\quad - 2\alpha_k \left(f(x^k) - f(\bar{x}) \right) + \alpha_k^2 \|d^k\|^2 \\
&\quad \vdots \\
&\leq \|x^1 - \bar{x}\|^2 - 2 \sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(\bar{x}) \right) + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2,
\end{aligned}$$

That is,

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^1 - \bar{x}\|^2 - 2\sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(\bar{x})\right) + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2. \tag{3.8}$$

As $\|x^{k+1} - \bar{x}\|^2 \geq 0$ and we are assuming that $\|x^1 - \bar{x}\| \leq R$, by the inequality (3.8), we obtain that

$$2\sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(\bar{x})\right) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2. \tag{3.9}$$

On the other hand, we have to

$$\begin{aligned}
\sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(\bar{x})\right) &\geq \sum_{i=1}^{k} \alpha_i \left(\min_{i=1,\ldots,k} \{f(x^i) - f(\bar{x})\}\right) \\
&= \min_{i=1,\ldots,k} \{f(x^i) - f(\bar{x})\} \sum_{i=1}^{k} \alpha_i \\
&= \left(f_{best}^k - f(\bar{x})\right) \sum_{i=1}^{k} \alpha_i,
\end{aligned}$$

thus,

$$2\sum_{i=1}^{k} \alpha_i \left(f(x^i) - f(\bar{x})\right) \geq 2\left(f_{best}^k - f(\bar{x})\right) \sum_{i=1}^{k} \alpha_i. \tag{3.10}$$

Now, combining the inequalities (3.9) and (3.10), we have

$$2\left(f_{best}^k - f(\bar{x})\right) \sum_{i=1}^{k} \alpha_i \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2,$$

therefore,

$$f_{best}^k - f(\bar{x}) \leq \frac{R^2 + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{3.11}$$

As we assume that $\|d^k\| \leq L$ for all k, then from the inequality (3.11), we obtain that

$$f_{best}^k - f(\bar{x}) \leq \frac{R^2 + L^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}. \tag{3.12}$$

From the inequality (3.12), we will obtain several convergence results.

Constant step size. Since $\alpha_k = \alpha$ for all k, then from inequality (3.12), we have

$$\begin{aligned}
f_{best}^k - f(\bar{x}) &\leq \frac{R^2 + L^2 \sum_{i=1}^{k} \alpha^2}{2\sum_{i=1}^{k} \alpha} \\
&= \frac{R^2 + L^2 k\alpha^2}{2k\alpha} \\
&= \frac{R^2}{2k\alpha} + \frac{L^2\alpha}{2},
\end{aligned}$$

therefore,

$$\lim_{k\to\infty} f_{best}^k - f(\bar{x}) \leq \frac{L^2\alpha}{2}.$$

Then, using the subgradient method with constant step size, f_{best}^k converges to a point in the interval $\left[f(\bar{x}), f(\bar{x}) + \frac{L^2\alpha}{2}\right]$. As a consequence, the precision depends on the step length value, as the smaller the value of α, the smaller $\frac{L^2\alpha}{2}$ will be.

Constant step length. Since $\alpha_k = \frac{\gamma}{\|d^k\|}$ for all k, then by the inequality (3.11) and by the fact that $\alpha_i = \frac{\gamma}{\|d^i\|} \geq \frac{\gamma}{L}$, we have

$$\begin{aligned}
f_{best}^k - f(\bar{x}) &\leq \frac{R^2 + \sum_{i=1}^k \alpha_i^2 \|d^i\|^2}{2\sum_{i=1}^k \alpha_i} \\
&= \frac{R^2 + \sum_{i=1}^k \frac{\gamma^2}{\|d^i\|^2}\|d^i\|^2}{2\sum_{i=1}^k \frac{\gamma}{\|d^i\|}} \\
&= \frac{R^2 + k\gamma^2}{2\sum_{i=1}^k \frac{\gamma}{\|d^i\|}} \\
&\leq \frac{R^2 + k\gamma^2}{2\sum_{i=1}^k \frac{\gamma}{L}} \\
&= \frac{R^2 + k\gamma^2}{2k\frac{\gamma}{L}} \\
&= \frac{LR^2}{2k\gamma} + \frac{L\gamma}{2},
\end{aligned}$$

therefore,

$$\lim_{k\to\infty} f_{best}^k - f(\bar{x}) \leq \frac{L\gamma}{2}.$$

Then, using the subgradient method with constant step length, f_{best}^k converges to a point in the interval $\left[f(\bar{x}), f(\bar{x}) + \frac{L\gamma}{2}\right]$. Consequently, the precision depends on the value of the step length, since the smaller the value of γ, the smaller $\frac{L\gamma}{2}$ will be.

Square summable but not summable. Since we choose the sequence $\{\alpha_k\}$ of step sizes such that it satisfies the following conditions:

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = +\infty, \quad \bar{\alpha} = \sum_{k=1}^{\infty} \alpha_k^2 < +\infty.$$

then by the inequality (3.12), we have that

$$\begin{aligned}
f_{best}^k - f(\bar{x}) &\leq \frac{R^2 + L^2\sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} \\
&\leq \frac{R^2 + L^2\sum_{i=1}^{\infty} \alpha_i^2}{2\sum_{i=1}^k \alpha_i} \\
&= \frac{R^2 + L^2\bar{\alpha}}{2\sum_{i=1}^k \alpha_i}.
\end{aligned}$$

Since $R^2 + L^2\bar{\alpha}$ is constant and $\lim_{k\to\infty}\left(\sum_{i=1}^{k}\alpha_i\right) = +\infty$ then $\lim_{k\to\infty}\frac{R^2+L^2\bar{\alpha}}{2\sum_{i=1}^{k}\alpha_i} = 0$, therefore,

$$\lim_{k\to\infty} f_{best}^{k} - f(\bar{x}) = 0.$$

Then, using the subgradient method with step length square summable but not summable, f_{best}^{k} converges to $f(\bar{x})$, which is the solution to the problem.

Nonsummable diminishing. We choose the sequence $\{\alpha_k\}$ of step sizes such that it satisfies the conditions:

$$\alpha_k \geq 0, \quad \lim_{k\to\infty}\alpha_k = 0, \quad \sum_{k=1}^{\infty}\alpha_k = +\infty,$$

and let $\varepsilon > 0$. Since $\lim_{k\to\infty}\alpha_k = 0$, there exists $n_1 \in \mathbb{N}$ such that if $i > n_1$, then $\alpha_i \leq \frac{\varepsilon}{L^2}$. On the other hand, since $\lim_{k\to\infty}\left(\sum_{i=1}^{k}\alpha_i\right) = +\infty$, there exists $n_2 \in \mathbb{N}$ such that

$$\sum_{i=1}^{n_2}\alpha_i \geq \frac{1}{\varepsilon}\left(R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2\right). \tag{3.13}$$

Let $n_0 = \max\{n_1, n_2\}$. Then for $k > n_0$, the right side of the inequality (3.12) can be written as

$$\begin{aligned}\frac{R^2 + L^2\sum_{i=1}^{k}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} &= \frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} + \frac{L^2\sum_{i=n_1+1}^{k}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} \\ &= \frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} + \frac{L^2\sum_{i=n_1+1}^{k}\alpha_i^2}{2\sum_{i=1}^{n_1}\alpha_i + 2\sum_{i=n_1+1}^{k}\alpha_i}.\end{aligned} \tag{3.14}$$

Now let us analyze each part of the relationship separately (3.14). For the first installment, since $n_2 < k$, we have

$$\frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} \leq \frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{n_2}\alpha_i}.$$

Using the relation (3.13), we have

$$\frac{1}{\sum_{i=1}^{n_2}\alpha_i} \leq \frac{\varepsilon}{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2},$$

therefore,

$$\begin{aligned}\frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{k}\alpha_i} &\leq \frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2\sum_{i=1}^{n_2}\alpha_i} \\ &\leq \frac{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}{2}\left(\frac{\varepsilon}{R^2 + L^2\sum_{i=1}^{n_1}\alpha_i^2}\right) \\ &= \frac{\varepsilon}{2}.\end{aligned}$$

For the second part of the relation (3.14), since $\alpha_i \geq 0$ for all i, we have

$$\frac{L^2 \sum_{i=n_1+1}^{k} \alpha_i^2}{2\sum_{i=1}^{n_1} \alpha_i + 2\sum_{i=n_1+1}^{k} \alpha_i} \leq \frac{L^2 \sum_{i=n_1+1}^{k} \alpha_i^2}{2\sum_{i=n_1+1}^{k} \alpha_i}.$$

Since $\alpha_i \leq \frac{\varepsilon}{L^2}$ if $i > n_1$, then $\alpha_i^2 \leq \alpha_i \frac{\varepsilon}{L^2}$ if $i > n_1$, thus, $\sum_{i=n_1+1}^{k} \alpha_i^2 \leq \sum_{i=n_1+1}^{k} \alpha_i \frac{\varepsilon}{L^2}$, and therefore

$$\frac{L^2 \sum_{i=n_1+1}^{k} \alpha_i^2}{2\sum_{i=n_1+1}^{k} \alpha_i} \leq \frac{L^2 \sum_{i=n_1+1}^{k} \alpha_i \frac{\varepsilon}{L^2}}{2\sum_{i=n_1+1}^{k} \alpha_i} = \frac{\varepsilon}{2}.$$

Finally, we conclude that

$$\begin{aligned}\frac{R^2 + L^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} &= \frac{R^2 + L^2 \sum_{i=1}^{n_1} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} + \frac{L^2 \sum_{i=n_1+1}^{k} \alpha_i^2}{2\sum_{i=1}^{n_1} \alpha_i + 2\sum_{i=n_1+1}^{k} \alpha_i} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,\end{aligned}$$

this shows that the right side of the inequality (3.12) converges to 0, and therefore, $\lim_{k\to\infty} f_{best}^k - f(\bar{x}) = 0$.

Then, using the subgradient method with nonsummable diminishing step size, f_{best}^k converges to $f(\bar{x})$, which is the solution to the problem.

Nonsummable diminishing step lengths. We choose the step size $\alpha_k = \frac{\gamma_k}{\|d^k\|}$ for all iterations, such that the sequence $\{\gamma_k\}$ satisfies the following conditions:

$$\gamma_k \geq 0, \quad \lim_{k\to\infty} \gamma_k = 0, \quad \sum_{k=1}^{\infty} \gamma_k = +\infty.$$

Using the inequality (3.11), we obtain that

$$\begin{aligned}f_{best}^k - f(\bar{x}) &\leq \frac{R^2 + \sum_{i=1}^{k} \alpha_i^2 \|d^i\|^2}{2\sum_{i=1}^{k} \alpha_i} \\ &= \frac{R^2 + \sum_{i=1}^{k} \gamma_i^2}{2\sum_{i=1}^{k} \frac{\gamma_i}{\|d^i\|}} \\ &\leq \frac{R^2 + \sum_{i=1}^{k} \gamma_i^2}{\frac{2}{L}\sum_{i=1}^{k} \gamma_i}.\end{aligned}$$

By the same analysis done in the case of nonsummable diminishing step size, it follows that $\frac{R^2+\sum_{i=1}^{k} \gamma_i^2}{\frac{2}{L}\sum_{i=1}^{k} \gamma_i}$ converges to 0, when $k \to \infty$, therefore $\lim_{k\to\infty} f_{best}^k - f(\bar{x}) = 0$.

Then, using the subgradient method with nonsummable diminishing step lengths, f_{best}^k converges to $f(\bar{x})$, which is the solution to the problem.

A bound on the suboptimality bound. In the inequality (3.12), we estimate the number $f_{best}^k - f(\bar{x})$ by $\frac{R^2+L^2\sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}$. Since $\frac{R^2+L^2\sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}$ is a convex and symmetric

function of $\alpha_1, \ldots, \alpha_k$, so it reaches its smallest value when $\alpha_i = \alpha$, for all $i = 1, \ldots, k$, see [4]. In this case, the optimal value is given by

$$\frac{R^2 + L^2 k\alpha^2}{2k\alpha}.$$

Since we have equality

$$\frac{R^2 + L^2 k\alpha^2}{2k\alpha} = \frac{\frac{R^2}{k\alpha} + L^2\alpha}{2},$$

using the inequality between the arithmetic and geometric means (A.M)≥(G.M) for the numbers $\frac{R^2}{k\alpha}$ and $L^2\alpha$, we obtain that

$$\begin{aligned}\frac{\frac{R^2}{k\alpha} + L^2\alpha}{2} &\geq \sqrt{\frac{R^2}{k\alpha} L^2\alpha} \\ &= \frac{RL}{\sqrt{k}},\end{aligned}$$

and equality occurs when

$$\frac{R^2}{k\alpha} = L^2\alpha,$$

that is, when

$$\alpha = \frac{R}{L\sqrt{k}} = \frac{(\frac{R}{L})}{\sqrt{k}}.$$

By this analysis, the choice of $\alpha_1, \ldots, \alpha_k$ that minimizes the estimate $\frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}$ is given by

$$\alpha_i = \frac{(\frac{R}{L})}{\sqrt{k}}, \quad i = 1, \ldots, k,$$

and with this choice we obtain that

$$f_{best}^k - f(\bar{x}) \leq \frac{RL}{\sqrt{k}}.$$

Therefore, if we made any other step size choice for $\alpha_1, \ldots, \alpha_k$, would have to

$$\frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i} \geq \frac{RL}{\sqrt{k}}.$$

Given $\varepsilon > 0$, so that the inequality

$$\frac{RL}{\sqrt{k}} < \varepsilon,$$

is true, it is necessary that

$$k > \left(\frac{RL}{\varepsilon}\right)^2.$$

this shows that if we use the estimate $\frac{R^2 + L^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}$ as the stopping criterion, then the number of steps needed to obtain a guaranteed accuracy of ε is at least $(RL/\varepsilon)^2$, for any choice of step size for $\alpha_1, \ldots, \alpha_k$. This shows that for this choice of stopping criterion, the subgradient method will be very slow.

Chapter 4

Subgradient method with non-monotone line search

The chapter will be divided into two sections. In the first section, we will present the subgradient projection method with non-monotone line search algorithm, proposed in [8]. This algorithm is not necessarily a descent method, but any potential increase in the function values is limited by a non-increasing sequence of parameters. Moreover, the step sizes are chosen adaptively using a non-monotone line search in the opposite direction to a subgradient. The main results include the well-definition of the algorithm and some inequalities that assist in the convergence proofs. In the second section, we will present the convergence analysis of the method under additional assumptions on the non-monotonicity sequence. The main results are the convergence proofs.

Consider the constrained minimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{C}, \end{aligned} \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\mathcal{C} \subset \mathbb{R}^n$ is a non-empty, convex and closed set. We denote the set of solutions to Problem (4.1) by Ω^* and the optimal value of the function f by f^*. In this chapter, we assume that:

(H1) $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and $L_{f,\mathcal{C}}$-Lipschitz continuous function;

(H2) $f^* := \inf_{x \in \mathcal{C}} f(x) > -\infty$.

4.1 The algorithm

Consider the following algorithm to solve the Problem (4.1):

Algorithm 4 SubGrad projection method with non-monotone line search

1: Fix $c > 0$, $(\gamma_k)_{k\in\mathbb{N}} \subset \mathbb{R}_{++}$ a non-increasing sequence, $\rho > 1/2$, $\beta \in (0,1)$ and $\alpha > 0$. Choose an initial point $x_1 \in C$. Set $\alpha_1 = \alpha$ and $k = 1$;

2: Choose $s_k \in \partial f(x_k)$. If $s_k = 0$, then STOP and return x_k;

3: Compute

$$l_k := \min\{l \in \mathbb{N} : \beta^l \alpha_k \leq c\beta\gamma_k, \quad f(\mathcal{P}_C(x_k - \beta^l \alpha_k s_k)) \leq f(x_k) - \rho(\beta^l \alpha_k)\|s_k\|^2 + \gamma_k\}; \tag{4.2}$$

4: Set $x_{k+1} := \mathcal{P}_C(x_k - \beta^{l_k}\alpha_k s_k)$, $\alpha_{k+1} := \beta^{l_k - 1}\alpha_k$. Update $k := k+1$ and return to Step 2.

Remark 4.1.1. *By the Proposition 1.0.8, the set where the function f is not differentiable is of zero measure, which means that f will almost always be differentiable. Since, in the case where the function f is differentiable at a point $x \in \mathbb{R}^n$ and convex, the Proposition 1.0.6 guarantees that $\partial f(x) = \{\nabla f(x)\}$, then almost every direction opposite a subgradient will be a descent direction. Thus, it is expected that the Algorithm 4 skips the points where the function f is not differentiable and which are not minima points, and behaves in a similar way to the gradient method with this non-monotone line search.*

The next lemma guarantees that it is possible to calculate l_k satisfying (4.2), as a consequence, we will obtain two inequalities that will be important throughout the chapter.

Lemma 4.1.1. *There exists l_k satisfying (4.2). As a consequence, the following inequalities hold:*

$$\alpha_{k+1} \leq c\gamma_k, \quad f(x_{k+1}) \leq f(x_k) - \rho\beta\alpha_{k+1}\|s_k\|^2 + \gamma_k, \quad \forall k \in \mathbb{N}, \tag{4.3}$$

and $x_{k+1} \in \mathcal{C}$ for all $k \in \mathbb{N}$.

Proof. Since the function f is continuous, the function $\mathcal{P}_\mathcal{C}$ is continuous, and the point $x_k \in \mathcal{C}$, we have that $\lim_{\alpha\to 0^+}(f(\mathcal{P}_\mathcal{C}(x_k - \alpha s_k)) - f(x_k) + \rho\alpha\|s_k\|^2) = 0$. Thus, given $\gamma_k > 0$, there exists $\eta_k > 0$ such that

$$f(\mathcal{P}_\mathcal{C}(x_k - \alpha s_k)) - f(x_k) + \rho\alpha\|s_k\|^2 < \gamma_k, \quad \forall \alpha \in (0, \eta_k],$$

or equivalently,

$$f(\mathcal{P}_{\mathcal{C}}(x_k - \alpha s_k)) \leq f(x_k) - \rho\alpha\|s_k\|^2 + \gamma_k, \quad \forall \alpha \in (0, \eta_k]. \tag{4.4}$$

Since $\beta \in (0,1)$, we have $\lim_{l\to\infty} \beta^l \alpha_k = 0$. Since $c\beta\gamma_k > 0$ there exists $\bar{l} \in \mathbb{N}$ such that $l \geq \bar{l}$ implies that $\beta^l \alpha_k \leq c\beta\gamma_k$. Since $\eta_k > 0$, there exists $\tilde{l} \in \mathbb{N}$ such that $l \geq \tilde{l}$ implies that $\beta^l \alpha_k \in (0, \eta_k]$. Taking $l^* = \max\{\bar{l}, \tilde{l}\}$, then $l \geq l^*$ implies that

$$\beta^l \alpha_k \leq c\beta\gamma_k, \quad f(\mathcal{P}_C(x_k - \beta^l \alpha_k s_k)) \leq f(x_k) - \rho(\beta^l \alpha_k)\|s_k\|^2 + \gamma_k.$$

Now it remains to take l_k as the smallest of the numbers $i = 0, 1, \ldots, l^*$ such that $\beta^i \alpha_k$ satisfies the two inequalities of (4.2), that is,

$$\beta^i \alpha_k \leq c\beta\gamma_k, \quad f(\mathcal{P}_C(x_k - \beta^i \alpha_k s_k)) \leq f(x_k) - \rho(\beta^i \alpha_k)\|s_k\|^2 + \gamma_k,$$

this proves that there is l_k satisfying (4.2).

From $\beta^{l_k}\alpha_k \leq c\beta\gamma_k$, we obtain that $\beta^{l_k-1}\alpha_k \leq c\gamma_k$, which by the definition of α_{k+1} in Step 4 means that $\alpha_{k+1} \leq c\gamma_k$.

From $f(\mathcal{P}_C(x_k - \beta^{l_k}\alpha_k s_k)) \leq f(x_k) - \rho(\beta^{l_k}\alpha_k)\|s_k\|^2 + \gamma_k$, from the definition of x_{k+1} and α_{k+1} in Step 4, we obtain that $f(x_{k+1}) \leq f(x_k) - \rho\beta\alpha_{k+1}\|s_k\|^2 + \gamma_k$.

From the definition of x_{k+1} and α_{k+1}, as $\mathcal{C}$ is convex and closed, it follows that $x_{k+1} \in \mathcal{C}$, for all k. □

In step 1 of the Algorithm 4, we fix the parameters that will be used during the iterations and the sequence $(\gamma_k)_{k\in\mathbb{N}}$ of non-monotonicity that will be used in the line search. In step 2, we calculate some subgradient $s_k \in \partial f(x_k)$, given that, by the Proposition 1.0.5 the set $\partial f(x_k)$ is non-empty. If $s_k = 0$, by the Proposition 1.0.7, we find the solution. In step 3, we calculate l_k satisfying (4.2). As we saw previously, the Lemma 4.1.1 guarantees the existence of l_k. Therefore, of the numbers $\beta^l\alpha_k$ that satisfy the two inequalities in (4.2), as $\beta \in (0,1)$, then $\beta^{l_k}\alpha_k$ is the biggest of them. In step 4, we use $\beta^{l_k}\alpha_k$ as the step size in the direction of $-s_k$, and project the point $x_k - \beta^{l_k}\alpha_k s_k$ in the set $\mathcal{C}$. Additionally, we use l_k to define $\alpha_{k+1} := \beta^{l_k-1}\alpha_k$ which will be used to compute the new step size. We do $k := k+1$ and repeat the procedure.

From the second inequality in (4.3), we have $f(x_{k+1}) \leq f(x_k) - \rho\beta\alpha_{k+1}\|s_k\|^2 + \gamma_k$ for all k. Since $-\rho\beta\alpha_{k+1}\|s_k\|^2 \leq 0$, it follows that

$$f(x_{k+1}) \leq f(x_k) + \gamma_k, \quad \forall k \in \mathbb{N}.$$

Since $\gamma_k > 0$ for all k, the inequality above shows that $f(x_k) < f(x_{k+1})$ can happen, since $-s_k$ may not be a direction of descent, but we certainly have that $f(x_{k+1}) \leq f(x_k) + \gamma_k$. Later, we will make hypotheses that cause γ_k to approach de 0 asymptotically.

Thus, the step size was chosen using a non-monotone line search in the opposite direction to the subgradient (made in calculating the number l_k) and the possible increase of the objective function is limited by a sequence of positive parameters that implicitly control step size as we saw in the previous paragraph.

In the next lemma, we will prove classical inequalities that will be useful in convergence proofs.

Lemma 4.1.2. *For every $x \in \mathbb{R}^n$, we have*

$$2\beta\alpha_{k+1}(f(x_k) - f(x)) \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \beta^2\alpha_{k+1}^2\|s_k\|^2, \quad \forall k \in \mathbb{N}. \tag{4.5}$$

Additionally, if f is a σ-strongly convex function, then

$$2\beta\alpha_{k+1}(f(x_k) - f(x)) \leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \beta^2\alpha_{k+1}^2\|s_k\|^2, \quad \forall k \in \mathbb{N}. \tag{4.6}$$

Proof. Since the inequality (4.6) reduces to inequality (4.5) when $\sigma = 0$, it suffices to prove (4.6). From the definition of x_{k+1} and α_{k+1} in Step 4 of Algorithm 4 and Proposition 1.0.13, we have

$$\begin{aligned}
\|x_{k+1} - x\|^2 &= \|\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k}\alpha_k s_k) - x\|^2 \\
&\leq \|x_k - \beta^{l_k}\alpha_k s_k - x\|^2 \\
&= \langle x_k - x - \beta^{l_k}\alpha_k s_k, x_k - x - \beta^{l_k}\alpha_k s_k\rangle \\
&= \|x_k - x\|^2 - 2\beta^{l_k}\alpha_k\langle s_k, x_k - x\rangle + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2 \\
&= \|x_k - x\|^2 + 2\beta\alpha_{k+1}\langle s_k, x - x_k\rangle + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2.
\end{aligned} \tag{4.7}$$

Now, since we are assuming that f is a σ-strongly convex function, by Proposition 1.0.9, we have $\langle s_k, x - x_k\rangle \leq f(x) - f(x_k) - (\sigma/2)\|x_k - x\|^2$. Combining this information with inequality (4.7) we obtain that

$$\begin{aligned}
\|x_{k+1} - x\|^2 &\leq \|x_k - x\|^2 + 2\beta\alpha_{k+1}\langle s_k, x - x_k\rangle + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2 \\
&\leq \|x_k - x\|^2 + 2\beta\alpha_{k+1}\Big(f(x) - f(x_k) - (\sigma/2)\|x_k - x\|^2\Big) + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2 \\
&= \|x_k - x\|^2 + 2\beta\alpha_{k+1}\left(f(x) - f(x_k)\right) - \sigma\beta\alpha_{k+1}\|x_k - x\|^2 + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2,
\end{aligned}$$

thus,

$$2\beta\alpha_{k+1}(f(x_k) - f(x)) \leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x\|^2 - \|x_{k+1} - x\|^2 + (\beta^{l_k})^2(\alpha_k)^2\|s_k\|^2,$$

or even,

$$2\beta\alpha_{k+1}(f(x_k) - f(x)) \leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \beta^2\alpha_{k+1}^2\|s_k\|^2.$$

This proves inequality (4.6). □

The next lemma shows a relationship between the sequences $(\alpha_k)_{k\in\mathbb{N}}$ and $(\gamma_k)_{k\in\mathbb{N}}$. It will be used to show an inequality that helps in proofs of convergence.

Lemma 4.1.3. *The following inequality occurs:*

$$\alpha_k \geq \min\left\{\alpha_1, c\beta\gamma_k, \frac{\gamma_k}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}, \quad \forall k \in \mathbb{N}. \tag{4.8}$$

Proof. For $k = 1$, it is clear that $\alpha_1 \geq \min\left\{\alpha_1, c\beta\gamma_1, \frac{\gamma_1}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}$. Suppose, by contradiction, that there exists $k \in \mathbb{N}$ such that

$$\alpha_{k+1} < \min\left\{\alpha_1, c\beta\gamma_{k+1}, \frac{\gamma_{k+1}}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}. \tag{4.9}$$

As we are assuming that the sequence $(\gamma_k)_{k\in\mathbb{N}}$ is non-increasing, by the definition of α_{k+1} in Step 4 of the algorithm (4) and by the inequality (4.9), we have

$$\beta^{l_k-1}\alpha_k = \alpha_{k+1} < \min\left\{\alpha_1, c\beta\gamma_{k+1}, \frac{\gamma_{k+1}}{(1+\rho)L_{f,\mathcal{C}}^2}\right\} \leq \min\left\{c\beta\gamma_k, \frac{\gamma_k}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}, \tag{4.10}$$

thus,

$$\beta^{l_k-1}\alpha_k \leq c\beta\gamma_k. \tag{4.11}$$

As a function f is $L_{f,\mathcal{C}}$-Lipschitz continuous and the point $x_k \in \mathcal{C}$, by Proposition (1.0.13), we have

$$\begin{aligned} f\big(\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k-1}\alpha_k s_k)\big) - f(x_k) &\leq L_{f,\mathcal{C}}\|\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k-1}\alpha_k s_k) - x_k\| \\ &\leq L_{f,\mathcal{C}}\|x_k - \beta^{l_k-1}\alpha_k s_k - x_k\| \\ &= L_{f,\mathcal{C}}\beta^{l_k-1}\alpha_k\|s_k\|. \end{aligned} \tag{4.12}$$

By Proposition 1.0.10, we have $\|s_k\| \leq L_{f,\mathcal{C}}$. Thus, combining this inequality with inequality (4.12), we obtain

$$\begin{aligned} f\left(\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k-1}\alpha_k s_k)\right) - f(x_k) + \rho\beta^{l_k-1}\alpha_k\|s_k\|^2 &\leq L_{f,\mathcal{C}}\beta^{l_k-1}\alpha_k\|s_k\| + \rho\beta^{l_k-1}\alpha_k\|s_k\|^2 \\ &\leq L_{f,\mathcal{C}}^2\beta^{l_k-1}\alpha_k + \rho\beta^{l_k-1}\alpha_k L_{f,\mathcal{C}}^2 \\ &= L_{f,\mathcal{C}}^2\left(\beta^{l_k-1}\alpha_k + \rho\beta^{l_k-1}\alpha_k\right) \\ &= L_{f,\mathcal{C}}^2\left(\beta^{l_k-1}\alpha_k\right)(1+\rho). \end{aligned} \tag{4.13}$$

Using inequality (4.10), we have

$$\beta^{l_k-1}\alpha_k < \frac{\gamma_k}{(1+\rho)L_{f,\mathcal{C}}^2},$$

which, combined with inequality (4.13) gives us

$$\begin{aligned} f\left(\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k-1}\alpha_k s_k)\right) - f(x_k) + \rho\beta^{l_k-1}\alpha_k\|s_k\|^2 &\leq L_{f,\mathcal{C}}^2\left(\beta^{l_k-1}\alpha_k\right)(1+\rho) \\ &< L_{f,\mathcal{C}}^2\left(\frac{\gamma_k}{(1+\rho)L_{f,\mathcal{C}}^2}\right)(1+\rho) \\ &= \gamma_k, \end{aligned}$$

thus,

$$f\left(\mathcal{P}_{\mathcal{C}}(x_k - \beta^{l_k-1}\alpha_k s_k)\right) < f(x_k) - \rho\beta^{l_k-1}\alpha_k\|s_k\|^2 + \gamma_k. \tag{4.14}$$

We conclude that inequalities (4.11) and (4.14) hold true, which contradicts the definition of l_k, since in this case, $l_k - 1$ satisfies both inequalities in (4.2). Therefore, inequality (4.8) holds for all k. □

Remark 4.1.2. *The choice of α_1 is crucial for the method's performance. In Ferreira et al. [8], there is no theoretically founded criterion for choosing α_1 in the context of line search methods. In cases where we know the constant $L_{f,C}$, a conservative choice is given by*

$$\alpha_1 = \min\left\{c\beta\gamma_1, \frac{\gamma_1}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}.$$

because in this case, it follows from the Lemma 4.1.3 that the line search condition is already satisfied with $l_1 = 0$.

In the next lemma, we will prove important inequalities that will be used in the convergence proofs of the sequence $(x_k)_{k\in\mathbb{N}}$. To do so, we will combine the inequalities (4.3) with Lemmas 4.1.2 and 4.1.3, and define the following positive constants for $\rho > 1/2$:

$$\Theta := \min\left\{\frac{\alpha_1}{\gamma_1}, c\beta, \frac{1}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}, \qquad \Gamma := \Theta\left(2\beta - \frac{\beta}{\rho}\right). \tag{4.15}$$

Lemma 4.1.4. *Suppose $\Omega^* \neq \varnothing$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 4 and let $x^* \in \Omega^*$. Then, the following inequality holds:*

$$\Gamma\gamma_{k+1}(f(x_k) - f^*) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2, \quad \forall k \in \mathbb{N}. \tag{4.16}$$

Additionally, if f is a σ-strongly convex function, we have:

$$\Gamma\gamma_{k+1}(f(x_k) - f^*) \leq (1 - \sigma\beta\Theta\gamma_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2, \quad \forall k \in \mathbb{N}. \tag{4.17}$$

Proof. Since the inequality (4.17) becomes the inequality (4.6) when $\sigma = 0$, then it is sufficient to prove the inequality (4.17). By Lemma 4.1.1, we have

$$\beta\alpha_{k+1}\|s_k\|^2 \leq \frac{f(x_k) - f(x_{k+1}) + \gamma_k}{\rho}.$$

On the other hand, taking x^* in the inequality (4.6) of the Lemma 4.1.2, we obtain that

$$2\beta\alpha_{k+1}(f(x_k) - f^*) \leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \beta^2\alpha_{k+1}^2\|s_k\|^2.$$

Now, let's see that

$$\begin{aligned}
\beta^2\alpha_{k+1}^2\|s_k\|^2 &= \beta\alpha_{k+1}\beta\alpha_{k+1}\|s_k\|^2 \\
&\leq \beta\alpha_{k+1}\left(\frac{f(x_k) - f(x_{k+1}) + \gamma_k}{\rho}\right) \\
&= \beta\alpha_{k+1}\left(\frac{f(x_k) - f(x_{k+1})}{\rho} + \frac{\gamma_k}{\rho}\right) \\
&= \beta\alpha_{k+1}\left(\frac{f(x_k) - f(x_{k+1})}{\rho}\right) + \frac{\beta\alpha_{k+1}\gamma_k}{\rho} \\
&\leq \beta\alpha_{k+1}\left(\frac{f(x_k) - f^*}{\rho}\right) + \frac{1}{\rho}\beta\alpha_{k+1}\gamma_k,
\end{aligned}$$

thus,

$$\begin{aligned}
2\beta\alpha_{k+1}(f(x_k) - f^*) &\leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \beta^2\alpha_{k+1}^2\|s_k\|^2 \\
&\leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \beta\alpha_{k+1}\left(\frac{f(x_k) - f^*}{\rho}\right) + \\
&\quad + \frac{1}{\rho}\beta\alpha_{k+1}\gamma_k,
\end{aligned}$$

therefore,

$$\left(2\beta - \frac{\beta}{\rho}\right)\alpha_{k+1}(f(x_k) - f^*) \leq (1 - \sigma\beta\alpha_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta\alpha_{k+1}\gamma_k. \tag{4.18}$$

On the other hand, using the Lemma 4.1.3, considering that $(\gamma_k)_{k\in\mathbb{N}}$ is a non-increasing sequence, and using the first equality in (4.15), we have

$$\begin{aligned}\alpha_{k+1} &\geq \min\left\{\alpha_1, c\beta\gamma_{k+1}, \frac{\gamma_{k+1}}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}\\ &= \min\left\{\frac{\alpha_1}{\gamma_{k+1}}, c\beta, \frac{1}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}\gamma_{k+1}\\ &\geq \min\left\{\frac{\alpha_1}{\gamma_1}, c\beta, \frac{1}{(1+\rho)L_{f,\mathcal{C}}^2}\right\}\gamma_{k+1}\\ &= \Theta\gamma_{k+1}. \end{aligned} \tag{4.19}$$

Furthermore, by Lemma 4.1.1 we have $\alpha_{k+1} \leq c\gamma_k$, which combined with (4.18) e (4.19) guarantees what

$$\left(2\beta - \frac{\beta}{\rho}\right)\Theta\gamma_{k+1}(f(x_k) - f^*) \leq (1 - \sigma\beta\Theta\gamma_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2,$$

therefore,

$$\Gamma\gamma_{k+1}(f(x_k) - f^*) \leq (1 - \sigma\beta\Theta\gamma_{k+1})\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2.$$

□

Remark 4.1.3. *Comparing the inequalities (4.5) and (4.6) with the inequalities (4.16) and (4.17) respectively, we conclude that the inequalities (4.16) and (4.17) transfer, to the sequence $(\gamma_k)_{k\in\mathbb{N}}$ of non-monotonicity, classical conditions that are imposed for the step size sequence that control the behavior of the sequence (x_k) generated by the classic subgradient method, as we saw in chapter 3. The algorithm 4 uses adaptive step sizes, which are obtained over the course of iterations. In the classic case, the step sizes were pre-fixed. Furthermore, for each sequence (γ_k) that we choose, from the Lemma 4.1.1 we obtain that $\alpha_{k+1} \leq c\gamma_k$, for all k; and in the inequality (4.19) in the Lemma 4.1.4, we obtain that $\alpha_{k+1} \geq \Theta\gamma_{k+1}$, for all k. Combining this information, we have that the Algorithm 4 chooses, using a non-monotone line search, the step size α_k satisfying*

$$\Theta\gamma_{k+1} \leq \alpha_{k+1} \leq c\gamma_k, \quad \forall k \in \mathbb{N}.$$

Thus, the method allows different choices for the sequence (γ_k) that controls non-monotonicity.

4.2 Convergence analysis

For the convergence analysis, we will analyze the behavior of the sequence $(x_k)_{k\in\mathbb{N}}$ under the hypotheses (H1), (H2) and two additional hypotheses. Additional hypotheses will be used separately and only when explicitly stated. The new hypotheses are:

(H3) The sequence of non-monotonicity parameters $(\gamma_k)_{k\in\mathbb{N}}$ satisfies

$$\lim_{N\to+\infty} \frac{\sum_{k=1}^{N} \gamma_k^2}{\sum_{k=1}^{N} \gamma_{k+1}} = 0.$$

(H4) The sequence of non-monotonicity parameters $(\gamma_k)_{k\in\mathbb{N}}$ satisfies

$$\lim_{N\to+\infty} \frac{\sum_{k=1}^{N} \gamma_k^2}{N\gamma_{N+1}} = 0.$$

Hypotheses are made about the behavior of the sequence $(\gamma_k)_{k\in\mathbb{N}}$ and with them we obtain the following convergence results.

Theorem 4.2.1. *Assume that $\Omega^* \neq \varnothing$. Let $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 4 with $\rho > 1/2$ and $x^* \in \Omega^*$. Then, for each fixed $N \in \mathbb{N}$, the following inequality hold:*

$$\min\left\{f(x_k) - f^* : \ k = 1, \dots, N\right\} \leq \frac{1}{\Gamma}\left(\|x_1 - x^*\|^2 + \beta\rho^{-1}c\sum_{k=1}^{N}\gamma_k^2\right)\frac{1}{\sum_{k=1}^{N}\gamma_{k+1}}. \quad (4.20)$$

Consequently, if (H3) holds, then $\lim_{N\to+\infty} \min\left\{f(x_k) - f^* : \ k = 1, \dots, N\right\} = 0.$

Proof. Let $k \leq N$. By the inequality (4.16) in Lemma 4.1.4, we have

$$\Gamma\gamma_{k+1}(f(x_k) - f^*) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2.$$

Thus,

$$\begin{aligned}\sum_{k=1}^{N}\Gamma\gamma_{k+1}(f(x_k) - f^*) &\leq \sum_{k=1}^{N}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \frac{1}{\rho}\beta c\gamma_k^2\right)\\ &= \|x_1 - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \sum_{k=1}^{N}\frac{1}{\rho}\beta c\gamma_k^2\\ &\leq \|x_1 - x^*\|^2 + \sum_{k=1}^{N}\frac{1}{\rho}\beta c\gamma_k^2. \quad (4.21)\end{aligned}$$

On the other hand,

$$\sum_{k=1}^{N}\Gamma\gamma_{k+1}(f(x_k) - f^*) \geq \Gamma \min_{k=1,\dots,N}\left\{f(x_k) - f^*\right\}\sum_{k=1}^{N}\gamma_{k+1}. \quad (4.22)$$

Combining the inequalities (4.21) and (4.22), we obtain that

$$\Gamma \min_{k=1,\dots,N} \{f(x_k) - f^*\} \sum_{k=1}^{N} \gamma_{k+1} \leq \|x_1 - x^*\|^2 + \sum_{k=1}^{N} \frac{1}{\rho}\beta c \gamma_k^2,$$

therefore,

$$\min_{k=1,\dots,N} \{f(x_k) - f^*\} \leq \frac{1}{\Gamma}\Big(\|x_1 - x^*\|^2 + \frac{1}{\rho}\beta c \sum_{k=1}^{N} \gamma_k^2\Big) \frac{1}{\sum_{k=1}^{N} \gamma_{k+1}},$$

and this proves (4.20). Now, assuming that (H3) is valid and using (4.20), we have

$$\min_{k=1,\dots,N} \{f(x_k) - f^*\} \leq \frac{\|x_1 - x^*\|^2}{\Gamma \sum_{k=1}^{N} \gamma_{k+1}} + \frac{\frac{1}{\rho}\beta c \sum_{k=1}^{N} \gamma_k^2}{\Gamma \sum_{k=1}^{N} \gamma_{k+1}},$$

and how (H3) implies that $\lim_{N\to\infty} \frac{1}{\sum_{k=1}^{N} \gamma_{k+1}} = 0$, we concluded that

$$\lim_{N\to\infty} \min_{k=1,\dots,N} \{f(x_k) - f^*\} = 0.$$

□

The Theorem 4.2.1 provides an inequality that together with hypothesis (H3) guarantees that the sequence $\min\{f(x_k) : \ k = 1, \dots, N\}$ converges to f^*, provided that $\Omega^* \neq \varnothing$. This means that convergence information is obtained on the functional values.

If we assume that the sequence $(\gamma_k)_{k\in\mathbb{N}}$ satisfies the following hypotheses:

(H5) $\sum_{k=1}^{+\infty} \gamma_k^2 < +\infty$,

(H6) $\sum_{k=1}^{+\infty} \gamma_k = +\infty$,

then the next theorem ensures that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to the solution of the Problem (4.1) when the set $\Omega^* \neq \varnothing$.

Remark 4.2.1. *If $(\gamma_k)_{k\in\mathbb{N}}$ satisfies (H5) and (H6), then $(\gamma_k)_{k\in\mathbb{N}}$ also satisfies (H3). The sequence $(\gamma_k)_{k\in\mathbb{N}}$ with $\gamma_k = 1/k$ satisfies (H5) and (H6).*

Theorem 4.2.2. *Let $(x_k)_{k\in\mathbb{N}}$ be generated by Algorithm 4 with $\rho > 1/2$. Assume that (H5) holds. If $\Omega^* \neq \varnothing$, then $(x_k)_{k\in\mathbb{N}}$ is bounded. Moreover, if (H6) hold, then $(x_k)_{k\in\mathbb{N}}$ converges to a solution of Problem (4.1).*

Proof. Let $x \in \Omega^*$. By inequality (4.16) in Lemma 4.1.4, for every k, we have

$$\Gamma\gamma_{k+1}(f(x_k) - f^*) \leq \|x_k - x\|^2 - \|x_{k+1} - x\|^2 + \frac{1}{\rho}\beta c \gamma_k^2,$$

that is,

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 - \Gamma\gamma_{k+1}(f(x_k) - f^*) + \frac{1}{\rho}\beta c\gamma_k^2.$$

As $x \in \Omega^*$, we have $f(x_k) - f^* \geq 0$ for all k. Therefore,

$$\begin{aligned}\|x_{k+1} - x\|^2 &\leq \|x_k - x\|^2 - \Gamma\gamma_{k+1}(f(x_k) - f^*) + \frac{1}{\rho}\beta c\gamma_k^2 \\ &\leq \|x_k - x\|^2 + \frac{1}{\rho}\beta c\gamma_k^2,\end{aligned}$$

that is,

$$\|x_{k+1} - x\|^2 \leq \|x_k - x\|^2 + \frac{1}{\rho}\beta c\gamma_k^2, \quad \forall k \in \mathbb{N}.$$

Since $x \in \Omega^*$ is arbitrary, and by (H5) the sequence $\left(\frac{1}{\rho}\beta c\gamma_k^2\right)_{k\in\mathbb{N}}$ is summable, then by the last inequality and Definition 1.0.9, we conclude that the sequence $(x_k)_{k\in\mathbb{N}}$ is quasi-Féjer convergent to the set Ω^*. Since $\Omega^* \neq \varnothing$, by item (i) of Proposition 1.0.14, we have that the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded. This proves the first assertion.

Now, let us define a subsequence $(x_{k_N})_{N\in\mathbb{N}}$ of $(x_k)_{k\in\mathbb{N}}$ such that

$$f(x_{k_N}) - f^* := \min_{k=1,\ldots,N}\left\{f(x_k) - f^*\right\}, \quad N \in \mathbb{N}.$$

Since the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded, then the subsequence $(x_{k_N})_{N\in\mathbb{N}}$ is also bounded. By the Bolzano-Weierstrass Theorem, there is a subsequence of $(x_{k_N})_{N\in\mathbb{N}}$ that is convergent, therefore, without loss of generality we will assume that the sequence $(x_{k_N})_{N\in\mathbb{N}}$ is convergent and consider that $\lim_{N\to\infty} x_{k_N} = \bar{x}$. As we are assuming that (H5) and (H6) are valid, then (H3) and (H6) are valid, therefore, using the Theorem 4.2.1 we obtain $\lim_{N\to+\infty}(f(x_{k_N}) - f^*) = 0$, that is, $\lim_{N\to+\infty} f(x_{k_N}) = f^*$. Since the function f is continuous and $\lim_{N\to\infty} x_{k_N} = \bar{x}$, we have $f(\bar{x}) = f^*$, therefore $\bar{x} \in \Omega^*$. Again, as the sequence $(x_k)_{k\in\mathbb{N}}$ is quasi-Féjer, through item (ii) of the Proposition 1.0.14 we conclude that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to $\bar{x}$. □

Chapter 5

Conclusion

In this work, we conducted a study of the classical gradient and subgradient methods, as well as a subgradient method with a non-monotone line search for Lipschitz convex functions. The gradient method, being a descent method, utilizes step sizes chosen via exact and inexact line search. In contrast, the subgradient method is not necessarily a descent method and employs pre-determined step sizes, which are not selected through line search. The subgradient method with a non-monotone line search adaptively chooses steps through a non-monotone line search mechanism similar to the Armijo rule.

In the gradient method, we studied convergence results in the following cases: when the function is differentiable and has a Lipschitz-continuous gradient; when the function is differentiable and has a continuous gradient; and when the function is convex, differentiable, and has a continuous gradient. In the first two cases, we concluded that if the sequence generated by the method has cluster points, then these are critical points of the problem. In the last case, when we add the hypothesis of convexity, we ensure that if the solution set is non-empty, the sequence generated by the method converges to the solution of the problem.

In the subgradient method, we studied convergence results for the following step size choices in the convex case: constant step size, constant step length, square summable but not summable, nonsummable diminishing, and nonsummable diminishing step lengths. We concluded that in the cases of constant step size and constant step length rules, f_{best}^k converges to an interval "close" to the solution $f(\bar{x})$. In the cases of square summable but not summable, nonsummable diminishing, and nonsummable diminishing step lengths rules, we ensure that f_{best}^k converges to the solution $f(\bar{x})$.

In the subgradient method with non-monotone line search, we study convergence results under hypotheses in the non-monotonicity sequence that are similar to the hypotheses made in the classical subgradient case. We conclude that under the hypothesis (H3) $\min_{k=1,\ldots,N}\left\{f(x_k)\right\}$ converges to the solution f^* and under the hypotheses (H5) and (H6) the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a solution to the problem.

As future work, we will consider the quasiconvex and Lipschitz case. In Cruz Neto et al. [6], the authors consider the subgradient method with square summable but not summable step sizes and the Armijo search for continuously differentiable, quasiconvex, and Lipschitz functions. To this end, the Plastria subdifferential is considered. It is defined as follows:

$$\partial^P f(x) = \{v \in \mathbb{R}^n \,:\, f(y) < f(x) \implies \langle v, y - x\rangle \leq f(y) - f(x)\}\,.$$

The reason is that if $f : \mathbb{R}^n \to \mathbb{R}$ is a quasiconvex, differentiable and Lipschitz function (with Lipschitz constant L), $x_0 \in \mathbb{R}^n$ is such that $\nabla f(x_0) \neq 0$, then

$$g := \frac{L\nabla f(x_0)}{||\nabla f(x_0)||} \in \partial^P f(x_0);$$

see [6, Corollary 6]. Note that the subdifferential in the convex context is a particular case of the Plastria subdifferential, i.e., $\partial f(x) \subset \partial^P f(x)$. In this sense, the work of Cruz Neto et al. [6], can be seen as a generalization of the classical subgradient method to the quasiconvex context.

Since Ferreira et al. [8] demonstrated that computationally, in the convex case, the subgradient method with non-monotone line search is more efficient than classical step sizes, we intend to propose a non-monotone version of the subgradient method for quasiconvex functions using the Plastria subdifferential, and thereby extending the results of Ferreira et al. [8] and obtaining a more efficient method than that proposed in Cruz Neto et al. [6]

Recently, Lara et al. [13] studied the subgradient method with square summable but not summable step sizes for strongly quasiconvex functions using the strong subdifferential defined as follows: Let $h : \mathbb{R}^n \to \mathbb{R}$ be a function, $\beta > 0$, $\gamma \geq 0$, and $K \subseteq \mathbb{R}^n$. Then the (β, γ, K)-strong subdifferential of h at $\bar{x} \in K$ is given by

$$\begin{aligned}\partial^K_{\beta,\gamma} h(\bar{x}) :=& \{\xi \in \mathbb{R}^n : \max\{h(y), h(\bar{x})\} \geq h(\bar{x}) + \frac{\lambda}{\beta}\langle \xi, y - \bar{x}\rangle \\ &+ \frac{\lambda}{2}\left(\gamma - \frac{\lambda}{\beta} - \lambda\gamma\right)||y - \bar{x}||^2,\, \forall\, y \in K, \forall\, \lambda \in [0,1]\}.\end{aligned}$$

Taking into account that if $K \subset \mathbb{R}^n$ is a closed and convex set, $h : \mathbb{R}^n \to \mathbb{R}$ lower semicontinuous and strongly quasiconvex on K with modulus $\gamma > 0$ and $\beta > 0$, then $\partial^K_{\beta,\gamma} h(\bar{x}) \neq \varnothing$ for every $\bar{x} \in K$; see [12, Corollary 38(a)].

Using this approach, the authors removed the Lipschitz continuity hypothesis used in Cruz Neto et al. [6]. In this context, we intend to propose a version of the subgradient method for strongly quasiconvex functions with non-monotone line search, thereby obtaining a computationally more efficient method than that proposed by Lara et al. [13].

Bibliography

[1] Beck, A. *First-order methods in optimization*, 1 ed. SIAM, 2017.

[2] Bertsekas, D. P. *Nonlinear programming*, 2nd ed. Athena Scientific, 1999.

[3] Boyd, S., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

[4] Boyd, S., Xiao, L., and Mutapcic, A. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter 2004*, 01 (2003).

[5] Combettes, P. L. Quasi-fejérian analysis of some optimization algorithms. In *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, D. Butnariu, Y. Censor, and S. Reich, Eds., vol. 8 of *Studies in Computational Mathematics*. Elsevier, 2001, pp. 115–152.

[6] Cruz Neto, J. X., Lopes, J. O., and Travaglia, M. V. Algorithms for quasiconvex minimization. *Optimization 60*, 8-9 (2011), 1105–1117.

[7] Ermol'ev, Y. M. Methods of solution of nonlinear extremal problems. *Cybernetics 2*, 4 (1966), 1–14.

[8] Ferreira, O. P., Grapiglia, G. N., Santos, E. M., and Souza, J. C. O. A subgradient method with non-monotone line search. *Computational Optimization and Applications 84*, 2 (2023), 397–420.

[9] Fitzpatrick, P. *Advanced calculus*, 2 ed., vol. 5. American Mathematical Soc., 2009.

[10] Izmailov, A., and Solodov, M. *Otimização, volume 1: Condições de otimalidade, elementos de análise convexa e de dualidade*, 3 ed. IMPA, 2014.

[11] Izmailov, A., and Solodov, M. *Otimização, volume 2: Métodos computacionais*, 3 ed. IMPA, 2018.

[12] Kabgani, A., and Lara, F. Strong subdifferentials: theory and applications in nonconvex optimization. *Journal of Global Optimization 84*, 2 (2022), 349–368.

[13] Lara, F., Marcavillaca, R., and Choque, J. A subgradient projection method for quasiconvex minimization. available at Research Square, 2024. PREPRINT (Version 1).

[14] Lima, E. L. *Curso de análise*, vol. 2. IMPA, 2018.

[15] Nemirovskij, A. S., and Yudin, D. B. Problem complexity and method efficiency in optimization.

[16] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, 1 ed. Applied Optimization 87. Springer US, 2004.

[17] Nocedal, J., and Wright, S. J. *Numerical optimization*. Springer, 1999.

[18] Polyak, B. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics 3*, 4 (1963), 864–878.

[19] Polyak, B. T. Introduction to optimization.

[20] Ribeiro, A. A., and Karas, E. W. *Otimização contínua: aspectos teóricos e computacionais*. Cengage Learning, 2013.

[21] Shor, N. Z. *Minimization Methods for Non-Differentiable Functions*, 1 ed. Springer Series in Computational Mathematics 3. Springer-Verlag Berlin Heidelberg, 1985.

[22] Urruty, J.-B. H., and Lemaréchal, C. *Convex analysis and minimization algorithms*. Springer-Verlag, 1993.