Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization

Olivier Devolder

Université catholique de Louvain (UCL)
ICTEAM research institute
Center for Operations Research and Econometrics (CORE)

Thesis submitted in partial fulfillment of the requirements for the degree of Docteur en Sciences de l'Ingénieur

Dissertation committee:
Prof. François Glineur (Université catholique de Louvain, Supervisor)
Prof. Yurii Nesterov (Université catholique de Louvain, Supervisor)
Prof. Frédéric Bonnans (Ecole Polytechnique-INRIA, France)
Prof. Coralia Cartis (University of Edinburgh, United Kingdom)
Prof. Boris Mordukhovich (Wayne State University, United States)
Prof. Paul Van Dooren (Université catholique de Louvain, President)

March, 2013

Acknowledgments

A PhD thesis is a long journey that cannot be led to success without the help and support of many people.

First of all, I would like to thank warmly my two PhD supervisors, François Glineur and Yurii Nesterov, for providing me such invaluable support, advice and opportunities during these last four years, for all the enthralling research meetings that we have had together, for all their insightful ideas that have so often opened new research possibilities and for the time and the energy that they have spent in the preparation, the fine-tuning and the submission of our different papers. Working with François and Yurii has been for me a great privilege that I wish to every young researcher interested by the world of convex optimization.

I am also very grateful to Arkadi Nemirovski for the unique opportunity of a research stay in his research team at the Georgia Institute of Technology, for all the highly rewarding discussions that we had together and for his precious advice that has clearly improved the content of this thesis.

Many thanks to the members of my accompanying and defense committees Frédéric Bonnans (École Polytechnique), Coralia Cartis (University of Edinburgh), Boris Mordukhovich (Wayne State University) and the chair Paul Van Dooren (UCL) for the time spent in the reading of this long thesis, for all their comments and suggestions that have dramatically improved the presentation of the thesis content and for their visits here in Louvain-la-neuve for the thesis confirmation, the private defense and the public defense.

I am also grateful to the people who have interacted with this work, have helped

me by all the scientific discussions we have had and have given me the opportunity to present these results at numerous international conferences. Among many others, I would like to thank in particular Peter Richtarik and Martin Takáč from the University of Edinburgh, Michel Baes and Stefan Richter from ETH Zurich, Moritz Diehl from the KULeuven and Ion Necoara from Politehnica University of Bucharest.

During this PhD thesis, I had the pleasure to meet, work and share time with so many people that I wish to thank here. In particular, I thank my current or former colleagues Gauthier, Joachim, Robert, Abdel, Alejandro, Rafael, Arnaud, Vladimir and of course Nico (!) at CORE ; Quentin, FX and Nicolas in Euler ; and Mariya, Fuxin and Le during my research stay at Gatech.

I also want to thank warmly the administrative staff in particular Catherine Germain, Sylvie Mauroy at CORE and Isabelle Hisette, Nathalie Ponet in INMA. Their kindness, availability and efficiency have been an important help for me, allowing me to focus on my research.

It would have been impossible to accomplish this long work without the constant support of my family and friends. Many thanks to all of you !

Last but certainly not least, I want to thank lovingly Audrey for her constant support, for helping me in my long reflection process concerning my career path while keeping me focused on the thesis when my motivation was at the lowest.

To all of you, thank you so much for your support !

This thesis has been financially supported by a F.R.S.- FNRS (Fund for Scientific Research) fellowship. My research stay at the Georgia Institute of Technology has been funded by an additional F.R.S.-FNRS grant. I have received also on some occasion support from the Belgian Programme on Interuniversity Attraction Poles (initiated by the Belgian Federal Science Policy Office). I am very grateful for all these opportunities, which have allowed me to focus on my research with a large degree of intellectual and financial freedom.

Finally, I am eternally grateful to my Alma Mater, the Université catholique de Louvain, for every thing, scientifically and humanly, this amazing university has provided to me all along the years.

Executive Summary

For many large-scale convex optimization problems, when requirements on the solution accuracy are not too high, first-order methods are methods of choice due to their cheap iteration cost and global convergence properties. However, the proven efficiency of a first-order method is classically based on some important assumptions:

- ◇ exact first-order information for the objective function is obtainable at each iteration
- ◇ the objective function satisfies specific properties such as smoothness or strong convexity
- ◇ projections on the feasible set can be computed efficiently at each iteration.

The goal of this thesis is to extend the analysis and scope of first-order methods of smooth convex optimization beyond this comfort zone. We review three challenging difficulties directly related to the above three assumptions: inexactness in the first-order information, lack of smoothness of the objective function and presence of linear constraints (that complicate projections on the feasible set).

1 Inexact first-order information

One of the main contributions of this thesis is an extended analysis of the effect of inexact first-order information on existing first-order methods of smooth convex optimization, together with the development of new methods exhibiting a better behavior with respect to errors.

We introduce the novel notion of (δ, L)-oracle. It can be seen as a generalization of the exact oracle of a smooth convex function that allows now an error δ in the first-order information. Such kind of oracle naturally appears when the first-order information of a smooth convex function is computed approximately, at a shifted point or when we solve approximately the subproblem defining a max-type smooth convex function in the context of smoothing techniques, Moreau-Yosida regularization and Augmented Lagrangians.

Our analysis of existing first-order methods of smooth convex optimization when used with such kind of inexact oracle reveals two very different behaviors: the classical Gradient Method (GM) is slow but robust, whereas the Fast Gradient Method (FGM) is fast but sensitive to errors. More specifically, error accumulation in the FGM implies that solution with small target accuracy are sometimes impossible to obtain, which then forces the use of the slow GM. In addition, we prove that this link between speed of convergence and sensitivity to errors is unavoidable: the faster a first-order method is, the worse its robustness must be. There is no hope to develop a perfect method which would be as fast as the FGM and as robust as the GM.

In reaction to this observation, we have developed a novel method, the Intermediate Gradient Method (IGM). This method can be seen as a smart hybrid between FGM and GM and can exhibit a whole range of intermediate behaviors. The main benefit is that the IGM is able to reach target accuracies unreachable by the FGM in a significantly smaller number of iterations compared to what is needed using the GM.

We also consider the situation of a stochastic oracle, where the first-order information is affected by stochastic noise. This situation is in some sense more favorable compared to the deterministic one (at least when the oracle is unbiased, otherwise the bias plays the role of a deterministic error). Taking into account the stochastic nature of the first-order information, we show that it is possible to modify the GM and FGM in a way that decreases the stochastic noise effect up to zero. This cannot be expected in the deterministic case for which worst case analysis gives an essentially adversarial flavor. Furthermore, the rate of decrease of the stochastic noise effect does not depend on the speed of the method: the faster convergence rate of the FGM does not prevent us to decrease the stochastic noise at the same optimal rate as the GM.

2 Lack of Smoothness

Surprisingly, the notion of (δ, L)-oracle can also be used to represent lack of smoothness in the objective function: the exact first-order information of a

nonsmooth or weakly smooth convex function can be seen as a particular case of (δ, L)-oracle. As a consequence, first-order methods initially developed for smooth convex problems can also be applied to functions with weaker levels of smoothness, resulting in some sense in universal first-order methods. This breaks the wall that seemed at first sight to exist between smooth and nonsmooth convex optimization. In particular, we show that the FGM, when used with well-chosen stepsizes (that depends on the level of smoothness of the objective function), can be seen as a universal optimal first-order method, that reaches optimal complexity in the smooth, weakly smooth and nonsmooth cases.

3 Linearly constrained problems

Finally, we consider a situation where projections on the feasible set cannot be computed efficiently, namely the case of a convex feasible set, possibly infinite-dimensional, constrained by a set of linear inequalities. This complication makes the usual first-order methods non applicable. A natural approach consists in dualizing the linear constraints, obtaining an unconstrained dual problem that can be solved using a first-order method. However, the dual function being typically non differentiable, we can only guarantee relatively slow convergence, both for the primal and dual problems.

We propose a new efficient approach, the double smoothing technique, in order to solve such kind of problems. Instead of applying a first-order method directly to the nonsmooth dual function, we modify this function to make it smooth and strongly convex. On the one hand, the smoothness of the objective function allows us to solve efficiently the smoothed dual problem with a Fast Gradient Method. On the other hand, the strong convexity property allows us the reconstruct, from this nearly optimal dual solution, a nearly optimal and nearly feasible primal solution with the same level of accuracy.

This dual approach is particularly efficient when the linear constraints can be seen as the only coupling constraints between the variables. In this case, after dualization of the linear constraints, the dual objective function typically becomes separable and therefore easy to compute.

Table of contents

1 Introduction **1**

I First-order methods with exact information **7**

2 First-order Methods in Convex Optimization **9**
2.1 Convex Optimization Problems 10
2.2 Numerical Methods for Convex Optimization Problems 17
2.3 Classes of Convex Problems 26
2.4 First-Order Methods in Smooth Convex Optimization 31
2.5 First-order Methods in Nonsmooth Convex Optimization 49
2.6 First-order Methods beyond their comfort zone 61

3 Double Smoothing Technique for Linearly Constrained Convex Optimization Problems **65**
3.1 Subgradient method vs Smoothing techniques 69
3.2 Problem formulation and dual approach 71
3.3 Examples of problems with separable structure 73
3.4 Double Smoothing Technique 74
3.5 Strong duality and norm of dual optimal solutions 79
3.6 Solving the primal-dual problem 82
3.7 Practical implementation 86
3.8 Application in Optimal Control 88
3.9 Comparison with the literature and conclusion 96

II First-order methods with inexact information **99**

4 First-Order Methods with Inexact Oracle: the smooth convex case **101**
4.1 The (δ, L)-oracle 104
4.2 Functions defined by an optimization subproblem 116

4.3 Gradient Methods with (δ, L)-oracle . . . 122
4.4 Fast Gradient Method with (δ, L)-oracle . . . 125
4.5 Comparison Classical and Fast Gradient Methods . . . 129
4.6 Comparison with other types of inexact oracle . . . 133
4.7 Application to functions with lack of smoothness . . . 135
4.8 First-order methods and error accumulation . . . 139

5 First-Order Methods with Inexact Oracle: the smooth strongly convex case 143
5.1 The (δ, L, μ)-oracle . . . 146
5.2 Examples of (δ, L, μ)-oracle . . . 148
5.3 Primal Gradient Method with (δ, L, μ)-oracle . . . 156
5.4 Dual Gradient Method with (δ, L, μ)-oracle . . . 158
5.5 Fast Gradient Method with (δ, L, μ)-oracle . . . 162
5.6 Application to weakly smooth uniformly convex functions . . . 178
5.7 Lower bound on error increase . . . 182

6 Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle 185
6.1 The Intermediate Gradient Method (IGM) . . . 189
6.2 Link with existing methods . . . 197
6.3 Optimal choice of the coefficients . . . 199
6.4 Switching policy for the coefficients . . . 204
6.5 Improvement compared with existing methods . . . 218
6.6 Power policy for the coefficients . . . 227
6.7 Numerical Illustration . . . 232

7 Stochastic First Order Methods in Smooth Convex Optimization 237
7.1 Smooth convex problem with stochastic oracle . . . 241
7.2 Stochastic Primal Gradient Method . . . 245
7.3 Stochastic estimate functions . . . 253
7.4 Stochastic Dual Gradient Method . . . 254
7.5 Stochastic Fast Gradient Method . . . 261
7.6 Probability of large deviation . . . 270
7.7 Postoptimization: Accuracy certificate . . . 275
7.8 Numerical Experiments . . . 278

8 Conclusion 287
8.1 Extended Summary . . . 287
8.2 Directions for Further Research . . . 294

Table of Methods 299

Chapter 1

Introduction

Making decisions in an optimal way is a natural human desire. In many situations, we are looking for the best solution (the solution that minimizes or maximizes a given criterion) satisfying some constraints:

- ◇ An investment banker tries to buy and sell assets in order to maximize his expected return while keeping the corresponding risk reasonable.
- ◇ An airline tries to allocate airplanes and crews most effectively, given a flight schedule and limitations on duty periods.
- ◇ A PNA device tries to find a feasible route from A to B that minimizes the total length or the total duration of the journey.
- ◇ A food product company tries to determine how many products to ship from each factory to each warehouse in order to minimize total shipping cost while not exceeding factory supplies and satisfying warehouse demands.
- ◇ An engineer in electronics tries to design electronic devices with maximum performance while satisfying some technical and budget constraints.
- ◇ An e-commerce company provides suggestions of new products to its current clients, that maximize the estimated 'likelihood' that these new products will meet the centers of interest of the client.

When the criterion and the constraints can be described quantitatively, such kind of problems can be translated into a mathematical form. The quantities that we want to choose in an optimal way are the decision variables $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, the criterion that we want to minimize (or maximize) is the objective function f and the constraints are described by the condition

$x \in Q$ where $Q \subseteq \mathbb{R}^n$ is a feasible set. Mathematically, an optimization problem can be written in the general form

$$\min_{x \in Q} f(x). \tag{1.1}$$

Let us note that an optimization problem can always be put under a minimization form w.l.o.g. since $\max_{x \in Q} f(x) = -\min_{x \in Q} -f(x)$.

Problem (1.1) is very general and encompasses very different types of objective functions and feasible sets. Even restricting to differentiable objective functions, solving such a problem globally, i.e. finding a point x in Q for which f is minimal, is in general very difficult. For a majority of problems, there is no hope to solve the problem at hand (i.e. to find a closed-form optimal solution) and, more annoyingly, there is also no hope to design a universal numerical method able to solve every optimization problem of the form (1.1) in a reasonable computation time. Indeed, for any numerical method, it is possible to construct problem instances (see section 1.1.3 in [58]), even with moderate dimension, that are intractable, i.e. that need huge computational time. In some sense, without additional well-chosen assumptions on the objective function and on the feasible set, the general global optimization problem (1.1) is unsolvable.

However, it is possible to define subclasses of problems for which much more can be expected, for which we can design specific efficient algorithms that can reliably solve even large optimization problems with good accuracy. The class of convex optimization problems is perhaps the best example of such good family of problems. A convex optimization problem is nothing else than a problem of the form (1.1) where f is a convex function and Q is a convex set. These problems exhibit some very good properties: any local optimal solution is also a global optimal solution, necessary and sufficient optimality conditions can be obtained and these properties allow us to design efficient numerical methods, able to solve large instances of such problems in reasonable computational time.

Furthermore, the class of convex problems is far from being only of theoretical interest, many practical optimization problems are convex or can be rewritten as convex problems. A huge literature exists about problems arising in control, signal processing, machine learning, statistics, finance or medical science that can be modeled as convex optimization problems and solved efficiently using the corresponding methods of convex optimization. For the reader interested in the applications of convex optimization, we refer to the introductory books [19, 9] and to more specialized references [18, 70, 76, 69].

In this thesis, we are interested in large-scale convex problems i.e. problems where the number of variables n is large enough so that we have to develop

numerical methods for which the number of iterations required to solve the problem (up to a given target accuracy) and the individual cost of each iteration depend only weakly on the problem size.

For solving large-scale convex optimization problems, first-order methods (i.e. numerical methods using only the value of the function and its gradient at each search point) are methods of choice in view of their cheap iteration costs. Indeed, the iteration cost of interior-point methods grows (too) quickly with the problem size (see for example [54, 9, 58]). Therefore, when the size of the problem increases significantly, interior-point methods become impracticable, a single iteration requires a huge amount of computation time. First-order methods, in spite of a slower (sublinear) convergence rate, present a significantly cheaper iteration cost.

The study and the development of first-order methods for convex optimization problems, in the smooth and nonsmooth cases, have attracted a lot of attention in the last decades (the Bibliography of this thesis illustrates this phenomenon in a non exhaustive way). The reasons for this interest lie in a growing demand for efficient numerical methods for very large-scale convex problems in various fields of engineering (for example in machine learning or for compressed sensing problems in signal processing) as well as a number of recent important breakthroughs. Two very good examples of such breakthroughs are the development of smoothing techniques ([59, 60, 61]) and of efficient gradient methods for composite function ([62, 4]), allowing in both cases to solve some types of nonsmooth problems more efficiently using their structure.

When the objective function is differentiable, the simplest first-order method to be considered is the classical gradient method also known as the Primal Gradient Method (PGM). However, it is well-known that the complexity of this method is not optimal ([55, 58]). In smooth convex optimization, optimal first-order methods have been developed for different classes of problems since 1983 ([56, 57, 59, 77]). These numerical schemes, also called Fast Gradient Methods (FGM) outperform theoretically and often in practice (see for example [4, 5, 6, 7, 79]) the classical gradient method.

However, the proven efficiency of these first-order methods is based on some important assumptions:

1. Standard analysis of first-order methods assumes availability of exact first-order information. Namely, the exact values of the function and its gradient must be available at each search point.
2. First-order methods are typically designed for a specific problem class, defined by a specific level of smoothness and a specific level of convexity of

the objective function. The convergence analysis of first-order methods is valid only when the objective function belongs to the appropriate problem class.

3. First-order methods are originally designed for unconstrained problems. They can be adapted relatively easily and without complexity modification to constrained problems provided that projections on the feasible set can be computed easily (it is the case for example when this set is a ball, a box or the positive orthant).

In this thesis, we consider different situations where such assumptions cannot be made:

1. only approximate first-order information is available
2. objective function is not as smooth as expected
3. additional constraints are present in the problem.

We study the consequences of these new situations and propose different approaches in order to remedy the newly encountered difficulties (modification of the existing methods, development of new methods and techniques).

This thesis consists of two parts.

In the first part, we assume that exact first-order information is available. This part contains two chapters. Chapter 2 is a survey about existing first-order methods in convex optimization and their behaviors when used with exact first-order information. It recalls basic notions and existing results needed for the thesis. Chapter 3 considers the situation where first-order information is exact but where the presence of linear constraints makes projections on the feasible set very expensive and therefore the classical first-order methods not efficient.

The second part of the thesis considers different situations where the available first-order information is inexact. We consider different kinds of errors (deterministic or stochastic) and different kinds of problems (smooth convex or smooth strongly convex objective functions):

- ◇ Deterministic inexact first-order information for smooth convex problems: Chapters 4 and 6.
- ◇ Deterministic inexact first-order information for smooth strongly convex problems: Chapter 5
- ◇ Stochastic inexact first-order information for smooth convex problems: Chapter 7.

Despite its length, an accelerated reading of the thesis is possible. The end of Chapter 2 lists the different questions that we try to answer in this thesis. At the beginning of each chapter, we review the relevant subset of these questions and provide summarized answers, emphasizing in this way the main results of the chapter. In the final Chapter 8, we propose an extended summary of the thesis content, with emphasis on the main messages, the core results and the links between chapters. We conclude the thesis by suggesting different directions for further research.

The reader can also find after the conclusion, a table of methods, emphasizing the new schemes that have been developed in this thesis.

Part I

First-order methods with exact information

Chapter 2

First-order Methods in Convex Optimization

Contents

2.1 Convex Optimization Problems **10**
2.1.1 Convex set 11
2.1.2 Convex function 12
2.1.3 Properties of convex optimization problems . . . 15
2.2 Numerical Methods for Convex Optimization Problems . **17**
2.2.1 Oracle associated with an optimization problem 17
2.2.2 Performance of a numerical method 18
2.2.3 Large-scale assumption and consequences . . . 22
2.2.4 Setup for a first-order method 24
2.3 Classes of Convex Problems **26**
2.3.1 Convexity assumption 26
2.3.2 Smoothness assumption 27
2.4 First-Order Methods in Smooth Convex Optimization . **31**
2.4.1 Optimal complexity 31
2.4.2 Primal Gradient Method (PGM) 33
2.4.3 The machinery of estimate functions 38
2.4.4 Dual Gradient Method (DGM) 40
2.4.5 Fast Gradient Method (FGM) 44
2.5 First-order Methods in Nonsmooth Convex Optimization **49**
2.5.1 Optimal complexity 50
2.5.2 Subgradient/ Mirror-descent methods 52
2.5.3 Looking inside the black-box 56
2.6 First-order Methods beyond their comfort zone **61**

2.1 Convex Optimization Problems

Optimization is an important field of applied mathematics. Given a finite-dimensional linear vector space E, a feasible set $Q \subset E$ and an objective function $f : Q \to \mathbb{R}$, we are looking for a point in Q where f is minimum. Mathematically, an optimization problem can be described as follows

$$\min_{x \in Q} f(x). \tag{2.1}$$

Remark 2.1. Space E is endowed with a norm $\|\cdot\|_E$ and E^*, the dual space of E, with the corresponding dual norm $\|g\|_E^* = \sup_{y \in E}\{\langle g, y\rangle : \|y\|_E \leq 1\}$ where $\langle ., .\rangle$ denotes the dual pairing. Most results in this thesis are proved for an arbitrary norm. In some cases, we need to restrict ourselves to a Euclidean norm, defined by a given arbitrary positive definite self-adjoint operator $B : E \to E^*$ as

$$\|h\|_E = \|h\|_2 = \langle Bh, h\rangle^{1/2} \quad \forall h \in E \tag{2.2}$$

$$\|s\|_E^* = \|s\|_2^* = \langle s, B^{-1}s\rangle^{1/2} \quad \forall s \in E^*. \tag{2.3}$$

For the optimization problem (2.1)

- ◇ The problem dimension is defined by the dimension of E.
- ◇ A feasible solution is any point x that belongs to the feasible set Q.
- ◇ A global optimal solution is a feasible solution x^* such that

$$f(x^*) \leq f(x), \quad \forall x \in Q.$$

- ◇ A local optimal solution is a feasible solution x^* for which there exists $r > 0$ such that

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*, r) \cap Q$$

  where $B(x^*, r) = \{x \in E | \|x - x^*\|_E < r\}$.

- ◇ The optimal value of the problem is defined as $f^* = \inf_{x \in Q} f(x)$.
- ◇ The optimal solution set is defined as $X^* = \{x^* \in Q | f(x^*) = f^*\}$ which is nothing else but the set containing all the global optimal solutions of problem (2.1).

In this thesis, we are interested in convex optimization problems i.e. problems of the form (2.1) where f is a convex function and Q is a convex set.
Before presenting in more details the properties of convex optimization problems and some examples of such problems, let us recall the notion of convex set and convex function:

2.1.1 Convex set

Definition 2.1. A set $Q \subset E$ is convex if for any $x, y \in Q$ and for any $\lambda \in [0,1]$,

$$\lambda x + (1-\lambda)y \in Q.$$

Geometrically, the convexity of Q means that, taking two arbitrary points x and y in Q, the segment between these two points must also lie in this set.
Let us now give some examples of typical convex sets (see for example [58, 19]):

- $\diamond$ $\emptyset$, $\mathbb{R}^n$, $\mathbb{R}^n_+$, $\mathbb{R}^n_{++}$, any linear vector space E
- $\diamond$ Half-space: $\{x \in E | \langle g, x\rangle \leq \beta\}$ where $g \in E^*$ and $\beta \in \mathbb{R}$
- $\diamond$ Balls: $B(\overline{x}, r) = \{x \in E | \, \|x - \overline{x}\|_E < r\}$ and $B[\overline{x}, r] = \{x \in E | \, \|x - \overline{x}\|_E \leq r\}$ for any norm $\|.\|_E$ on E
- $\diamond$ Ellipsoid: $\mathcal{E} = \{x \in E | \langle A(x - \overline{x}), (x - \overline{x})\rangle \leq \beta\}$ where $\overline{x} \in E$, $\beta \in \mathbb{R}_+$ and $A : E \to E^*$ is a positive definite operator
- $\diamond$ Lorentz cone: $\mathbb{L}_n = \{x \in \mathbb{R}^n | \sqrt{(x^{(2)})^2 + ... + (x^{(n)})^2} \leq x^{(1)}\}$
- $\diamond$ The positive semidefinite cone: $S_n^+ = \{X \in S_n | X \succeq 0\}$ where S_n is the space of symmetric $n \times n$ matrices
- $\diamond$ Unit Simplex: $\Delta_n = \{x \in \mathbb{R}^n_+ | \sum_{i=1}^n x^{(i)} = 1\}$
- $\diamond$ Unit Spectahedron: $\sum_n^+ = \{X \in S_n | X \succeq 0, \text{Trace}(X) = 1\}$

In addition to these basic convex sets, it is easy to construct more sophisticated convex sets using mathematical operations that preserve convexity. If $Q_1 \subset E$ and $Q_2 \subset F$ are two convex sets and $\mathcal{A} : E \to F$ is a linear operator then (see [58])

- $\diamond$ the intersection $Q_1 \cap Q_2$ is convex (with $F = E$)
- $\diamond$ the Minkowski sum $Q_1 + Q_2$ is convex (with $F = E$)
- $\diamond$ the Cartesian product $Q_1 \times Q_2$ is convex
- $\diamond$ the conic hull $\mathcal{K}(Q_1) = \{z \in E | z = \beta x, x \in Q_1, \beta \geq 0\}$ is convex
- $\diamond$ the convex hull $Conv(Q_1, Q_2) = \{z \in E | z = \lambda x + (1-\lambda)y, x \in Q_1, y \in Q_2, \lambda \in [0,1]\}$ is convex (with $E = F$)
- $\diamond$ the affine image $\mathcal{A}(Q_1) = \{z \in F | z = \mathcal{A}(x), x \in Q_1\}$ is convex
- $\diamond$ the inverse affine image $\mathcal{A}^{-1}(Q_2) = \{z \in E | \mathcal{A}(z) \in Q_2\}$ is convex.

2.1.2 Convex function

Convexity

Denote by $\mathrm{dom}\, f$ the domain of function f.

Definition 2.2. A function $f : \mathrm{dom}\, f \to \mathbb{R}$ is convex if its domain $\mathrm{dom}\, f$ is convex and for all $x, y \in \mathrm{dom}\, f$ and $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

Geometrically, the convexity of f means that the straight line linking the points $(x, f(x))$ and $(y, f(y))$ is above the graph of f on the interval $[x, y]$.

Remark 2.2. By convexity of $\mathrm{dom}\, f$, we have that $\lambda x + (1-\lambda)y \in \mathrm{dom}\, f$ and f is therefore well-defined at this point.

Remark 2.3. When $-f$ is convex, we say that the function f is concave.

Different equivalent definitions of the convexity of f can be given:

Proposition 1. *(see for example Theorem 3.1.2 in [58]) A function $f : \mathrm{dom}\, f \to \mathbb{R}$ is convex on $\mathrm{dom}\, f$ iff its epigraph*

$$\mathrm{epi}\, f = \{(x, t) \in \mathrm{dom}\, f \times \mathbb{R} | f(x) \leq t\}$$

is a convex set.

Proposition 2. *(see for example Section 3.1.3 in [19]) Assume that $\mathrm{dom}\, f$ is convex and that f is differentiable (i.e. that its gradient ∇f exists at each point in $\mathrm{dom}\, f$, which is open). Then f is convex iff for all $x, y \in \mathrm{dom}\, f$*

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

Geometrically, this equivalent definition means that a differentiable function is convex iff all its linear approximations (i.e. its first-order Taylor expansions) are below the graph of f.

Proposition 3. *(see for example Section 3.1.4 in [19]) Assume that $\mathrm{dom}\, f$ is convex and that f is twice differentiable (i.e. that its Hessian $\nabla^2 f$ exists at each point in $\mathrm{dom}\, f$, which is open). Then f is convex iff for all $x \in \mathrm{dom}\, f$*

$$\nabla^2 f(x) \succeq 0.$$

A twice differentiable function is convex iff its Hessians only have non negative eigenvalues.

Let us give some examples of simple convex functions (see [19, 58]):

- ◇ Linear function: $f(x) = \langle g, x\rangle + \beta$ for any $g \in E^*$, any $\beta \in \mathbb{R}$
- ◇ Quadratic function with positive semidefinite operator: $f(x) = \langle Ax, x\rangle$ where $A : E \to E^*$ is positive semidefinite
- ◇ Norm function: $f(x) = \|x\|_E$
- ◇ Exponential: $f(x) = e^x$ (with $E = \mathbb{R}$)
- ◇ Power function: $f(x) = |x|^p$ (with $E = \mathbb{R}$ and $p \geq 1$)

and of simple mathematical operations that preserve convexity (see [19, 58]):

- ◇ Nonnegative Multiple: If f is convex then αf is also convex for any $\alpha \geq 0$
- ◇ Sum: If f_1 and f_2 are convex then $f_1 + f_2$ is also convex
- ◇ Composition with affine operator: If f is convex and $\mathcal{A}$ is an affine operator then $f(\mathcal{A}(x))$ is convex
- ◇ Pointwise Supremum: If $F(x, y)$ is convex in x for all $y \in Y$ (where Y is an arbitrary set) then $f(x) = \sup_{y\in Y} F(x, y)$ is convex.

Strong Convexity

In some cases, assuming that f is convex is not sufficient and we have to consider a stronger notion:

Definition 2.3. A function $f : \operatorname{dom} f \to \mathbb{R}$ is strongly convex if its domain $\operatorname{dom} f$ is convex and if there exists a constant $\mu > 0$ such that for all $x, y \in \operatorname{dom} f$ and $\lambda \in [0, 1]$, we have

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

For a strongly convex function, μ is the parameter of strong convexity.
When the function is differentiable or twice differentiable, the following equivalent definitions can be given

Proposition 4. *(Section 2.1.3 in [58, 33]) Assume that* $\operatorname{dom} f$ *is convex and that* f *is differentiable on its domain. Then* f *is strongly convex with parameter* $\mu > 0$ *iff for all* $x, y \in \operatorname{dom} f$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|x - y\|_E^2 .$$

Proposition 5. *(Section 2.1.3 in [58]) Assume that* $\mathrm{dom}\, f$ *is convex and that* f *is twice differentiable on its domain. Then* f *is strongly convex with parameter* $\mu > 0$ *iff for all* $x \in \mathrm{dom}\, f$

$$\langle \nabla^2 f(x)h, h\rangle \geq \mu \|h\|_E^2, \quad \forall h \in E.$$

In particular if $\|.\|_E$ *is the usual Euclidean norm of* $\mathbb{R}^n$*, the strong convexity of* f *is equivalent with the condition*

$$\nabla^2 f(x) \succeq \mu I_n$$

for all $x \in \mathrm{dom}\, f$.

Of course, any strongly convex function is also convex and a convex function can be seen as a strongly convex one but with parameter $\mu = 0$. Contrarily to the notion of convexity, the strong convexity, more precisely the parameter of strong convexity, also depends, on the choice of norm $\|.\|_E$.

Let us now give some examples (see [58, 36]) of strongly convex functions (we assume here that $E = \mathbb{R}^n$) and an important property of strongly convex functions

- $f(x) = \frac{1}{2}\|x\|_2^2$ is strongly convex with parameter $\mu = 1$ with respect to any Euclidean norm $\|.\|_2$
- $f(x) = \langle Ax, x\rangle$ where $A \succeq \mu I_n$ is strongly convex with parameter μ with respect to the usual Euclidean norm $\|x\|_2 = \langle x, x\rangle^{1/2}$
- $f(x) = \ln(n) + \sum_{i=1}^n x^{(i)} \ln(x^{(i)})$ is strongly convex with parameter $\mu = 1$ with respect to the l_1 norm $\|x\|_1 = \sum_{i=1}^n |x^{(i)}|$

Proposition 6. *(Lemma 2.1.4 in [58]) Assume that* f_i *is strongly convex with respect to* $\|.\|_E$ *with parameter* $\mu_i \geq 0$ *(*$i = 1, 2$*) then* $f_1 + f_2$ *is strongly convex with respect to* $\|.\|_E$ *with parameter* $\mu_1 + \mu_2$.

In particular, the sum of a strongly convex function, with parameter μ, with a convex function (i.e. with parameter $\mu = 0$) is strongly convex with parameter μ.

Subgradients of a nonsmooth convex function

For a differentiable function f, the convexity can be characterized by the inequality

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad \forall x, y \in \mathrm{dom}\, f.$$

This inequality, satisfied by the gradient of a convex function, leads to the notion of subgradients that generalizes the gradient for nondifferentiable convex function.

Definition 2.4. Let f be a convex function. A vector $g \in E^*$ is called a subgradient of f at point $x \in \operatorname{dom} f$ if for any $y \in \operatorname{dom} f$, we have

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

The set of all subgradients of f at x, $\partial f(x)$ is called the subdifferential of function f at point x.

For a differentiable function, the subdifferential $\partial f(x)$ contains only one element $g(x) = \nabla f(x)$. But this definition makes sense also for nondifferentiable functions. For example, the absolute value function $f(x) = |x|$ is such that $\partial f(0) = [-1, 1]$, $\partial f(x) = \{-1\}$ for all $x < 0$ and $\partial f(x) = \{1\}$ for all $x > 0$. The following result guarantees the existence of subgradients for a closed convex function (a convex function being closed if its epigraph is a closed set):

Proposition 7. *(see Theorem 3.1.13 in [58]) Let f be closed and convex and $x \in$ int $\operatorname{dom} f$. Then $\partial f(x)$ is a nonempty bounded set.*

2.1.3 Properties of convex optimization problems

Let us look at the properties of a convex problem i.e. an optimization problem $\min_{x \in Q} f(x)$ where both the objective function f and the feasible set Q are convex.

The most important property of convex optimization problems is certainly the global optimality of any local optimal solution

Theorem 2.1. *(see for example Proposition 3.1.1 in [10]) Let f be a convex function on a convex set Q and let $x^* \in Q \cap \operatorname{dom} f$ be a local minimizer of f on Q, then x^* is a global minimizer of f on Q. Moreover, the set X^* of all minimizers of f on Q is convex.*

For strongly convex functions, we can also guarantee that the optimization problem (2.1) has at most one optimal solution

Theorem 2.2. *(see Proposition 3.1.1 in [10]) If f is strongly convex, then the optimal solution set X^* is either empty or a singleton.*

Concerning the existence of an optimal solution, we have the following result:

Theorem 2.3. *(see Proposition 3.2.1 in [10]) Assume that f is convex, that Q is closed and convex and that for some $\gamma \in \mathbb{R}$, the set $\{x \in Q | f(x) \leq \gamma\}$ is nonempty and compact, then the optimal solution set X^* is nonempty, compact, and convex.*

In particular if f is a differentiable strongly convex function and Q is a closed convex set, this condition is satisfied (see Theorem 2.2.6 in [58]) and we can guarantee the existence and uniqueness of an optimal solution.

Another crucial property of convex problems is the existence of simple, necessary and sufficient optimality conditions. Let us start with the easiest case when f is differentiable and Q is the whole space E:

Theorem 2.4. *(Theorem 1.2.1 and 2.1.1. in [58]) Assume that $Q = E$ and that f is convex and differentiable. Then $x^* \in E$ is a (global) optimal solution for the problem $\min_{x \in E} f(x)$ iff*

$$\nabla f(x^*) = 0. \tag{2.4}$$

Remark 2.4. In the nonconvex case, $\nabla f(x^*) = 0$ is only a necessary condition for being a local optimal solution. Thanks to the convexity, this condition is also sufficient and guarantees global optimality.

When the problem is constrained, the condition $\nabla f(x^*) = 0$ is no longer valid. The necessary and sufficient optimality condition becomes:

Theorem 2.5. *(Theorem 2.2.5 in [58]) Let f be a differentiable convex function and Q be a closed convex set. Then $x^* \in Q$ is a (global) optimal solution for the problem $\min_{x \in Q} f(x)$ iff*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in Q. \tag{2.5}$$

Remark 2.5. When $Q = E$, (2.5) is equivalent with (2.4).

For nondifferentiable problems, we cannot use the gradient in the optimality condition, but equivalent results can be obtained replacing the gradient by an arbitrary subgradient. More precisely, we have

Theorem 2.6. *(Theorem 3.1.15 in [58]) Assume that $Q = E$ and that f is convex. Then $x^* \in E$ is a (global) optimal solution for the problem $\min_{x \in E} f(x)$ iff*

$$0 \in \partial f(x^*) \tag{2.6}$$

and

Theorem 2.7. *(Theorem 5.4.7 in [10]) Let f be a convex function and Q be a closed convex set. Then $x^* \in Q$ is a (global) optimal solution for the problem $\min_{x \in Q} f(x)$ iff there exists $g \in \partial f(x^*)$ such that*

$$\langle g, x - x^* \rangle \geq 0, \quad \forall x \in Q. \tag{2.7}$$

2.2 Numerical Methods for Convex Optimization Problems

Few optimization problems can be solved at hand (i.e. have a closed-form optimal solution). In order to solve an optimization problem (the exact meaning of solving an optimization problem will be explained carefully in subsection 2.2.2), we have to use a numerical method. Its goal is to generate a sequence of approximate solutions that converges to an optimal solution.

Let us now explain what kind of information can be used by a numerical method in order to solve the problem, what is the generic form of an optimization numerical method and how we can measure the efficiency of a particular method.

2.2.1 Oracle associated with an optimization problem

When we want to solve a particular convex optimization problem using a numerical method, this problem is of course characterized by the convex function f that we want to minimize, the convex set Q on which we want to minimize f but also by the kind of information about f that can be used in the numerical method for solving this problem.

At each iteration k of a numerical method, using the current search point $x_k \in Q$, the method collects some new information about the problem and uses it (sometimes with the previously accumulated information) in order to construct the new search point $x_{k+1} \in Q$.

This process of collecting information about the problem is typically described using the notion of an oracle.

Definition 2.5. An oracle $\mathcal{O}$ is a unit that computes for the numerical method, at each call, information about the problem.

An optimization problem $\mathcal{P}$ is therefore characterized by the triplet (f, Q, $\mathcal{O}$). A standard assumption for the oracle is that it is black box and local i.e.

1. **Black-box**: The only information available for the numerical method is the answers of the oracle. The method has no access to the particular structure of the objective function f.

2. **Local**: A small variation far enough from the current search point x_k does not change the answer.

For most of the problems considered in the thesis, the local black-box assumption is satisfied. However, we will also see some situations where it is preferable

to use the explicit structure of the objective function (see section 2.5.3 and Chapter 3).

Typical oracles are black-box units that, given a search point x_k, compute at this point the value of the function and of its derivatives up to a given order. Depending on the maximum order of derivative computed, we can define different kinds of oracle:

1. A zero-order oracle is a unit that, given a search point x_k, computes the value of the function at this point: $\mathcal{O}_0(x_k) = f(x_k)$.

2. If the function is differentiable, a first-order oracle is a unit that, given a search point x_k, computes the value of the function and the gradient at this point: $\mathcal{O}_1(x_k) = (f(x_k), \nabla f(x_k))$. For a non differentiable convex function, a first-order oracle computes the value of the function and an arbitrary subgradient at the search point: $\mathcal{O}_1(x_k) = (f(x_k), g(x_k))$ where $g(x_k) \in \partial f(x_k)$.

3. If the function is twice differentiable, a second-order oracle is a unit that, given a search point x_k, computes the value of the function, the gradient and the Hessian of f at x_k: $\mathcal{O}_2(x_k) = (f(x_k), \nabla f(x_k), \nabla^2 f(x_k))$.

Related to the kind of oracle used by the scheme, we can define different families of numerical methods.

A (black-box) zero-order method is a method that uses, as only information about the particular problem instance, the answers of a zero-order oracle i.e. the value of the objective function at some search points.

Similarly, a (black-box) first-order method is a method based on the answers of a first-order oracle. The only information used by the method about the problem is the value of f and one of its subgradients at the search points.
More generally, a first-order method (not especially black-box) is a method that performs computations based on the function value and its (sub)gradient(s) (not the Hessian).

Similar definitions can be given for second-order methods (black-box or not).

2.2.2 Performance of a numerical method

When we want to solve an optimization problem $\mathcal{P} = (f, Q, \mathcal{O})$ using a numerical scheme, it is typically hopeless to expect obtaining an exact optimal solution x^* (i.e. such that $f(x^*) = f^*$) after a finite number of iterations.

A more realistic goal is to obtain an approximate solution. More precisely, given a target accuracy $\epsilon > 0$ and an optimality measure $Opt(x)$, we want to obtain after a finite number of iterations an ϵ-solution i.e. an approximate solution $y_k \in Q$ such that $Opt(y_k) \leq \epsilon$.

Different optimality measures can be considered. We can measure the accuracy of a feasible solution $x \in Q$

- ◇ by the non-optimality gap in term of the objective function $Opt_f(x) = f(x) - f^*$
- ◇ by the distance to the optimal solution set $Opt_d(x) = dist(x, X^*)$ where $X^* = \{x^* \in Q | f(x^*) = f^*\}$ or
- ◇ if f is differentiable, by the norm of the gradient $Opt_g(x) = \|\nabla f(x)\|_E^*$.

For every optimality measure, we have clearly $Opt(x^*) = 0$ for any $x^* \in X^*$ (assuming however that $Q = E$ for Opt_g).

Typical black-box numerical methods update (at least) two sequences of points:

1. a sequence of search points $\{x_k\}_{k\geq 0}$ where the oracle is called, i.e. where the information is computed
2. a sequence of approximate solution $\{y_k\}_{k\geq 0}$ for which an accuracy guarantee can be certified i.e. for which an upper-bound on $Opt(y_k)$ can be given.

Remark 2.6. The two sequences can be equal, i.e. $x_k = y_k$, for some numerical methods.

Every black-box method $\mathcal{M}$ considered in this thesis can therefore be put under a generic form. Let k be the iteration counter of the method and $\mathcal{I}_k$ the information set after k iterations that contains all the information collected by the oracle over the problem during the k first iterations.

Algorithm 1 Generic Numerical Scheme

Choose $x_0 \in Q$, independently of the problem instance.
Let $\mathcal{I}_{-1} = \emptyset$
for k = 0 : ... **do**
  Call oracle at x_k, obtaining $\mathcal{O}(x_k)$
  Update the information set: $\mathcal{I}_k = \mathcal{I}_{k-1} \cup (x_k, \mathcal{O}(x_k))$
  Using $\mathcal{I}_k$, generate a new approximate solution y_k
  Using $\mathcal{I}_k$, generate a new search point x_{k+1}
end for

In this generic scheme, the number of iterations that we have to perform is not specified. Two different approaches can be considered:

- ◇ Checking at the end of each iteration if $Opt(y_k) \leq \epsilon$ i.e. if the desired target accuracy is reached. However, such stopping criterion can be checked only if the optimality measure $Opt(y_k)$ can be computed in practice without the knowledge of the optimal solution set. This is possible for $Opt_g(.)$ but not for $Opt_f(.)$ nor $Opt_d(.)$. Furthermore, without an extra analysis of the scheme, we have no idea a priori of the needed number of iteration for reaching the stopping criterion.

- ◇ Establishing an a priori upper-bound K on the number of iterations after which we have the guarantee to obtain a solution with accuracy ϵ (i.e. such that $Opt(y_k) \leq \epsilon$).

This latter approach requires that we study theoretically the behavior of the method $\mathcal{M}$ on the problem $\mathcal{P}$. This leads us to the related notions of convergence rate and complexity of a particular method $\mathcal{M}$ on a particular problem $\mathcal{P}$.

Definition 2.6. Given an optimality measure $Opt(.)$, the convergence rate of the method $\mathcal{M}$ applied to the problem $\mathcal{P} = (f, Q, \mathcal{O})$ is defined by the function $\mathbb{N} \to \mathbb{R} : k \to ConvRate(\mathcal{P}, \mathcal{M}, k) := Opt(y_k)$(where $\{y_k\}_{k\geq 0}$ is the sequence of approximate solutions generated by $\mathcal{M}$ on $\mathcal{P}$).

Definition 2.7. Given an optimality measure $Opt(.)$ and a target accuracy $\epsilon > 0$, the complexity of the method $\mathcal{M}$ on the problem $\mathcal{P}$: $Compl(\mathcal{P}, \mathcal{M}, \epsilon)$ is defined as the number of iterations needed for solving the problem $\mathcal{P}$ up to accuracy ϵ i.e. for finding an iterate y_k such that $ConvRate(\mathcal{P}, \mathcal{M}, k) \leq \epsilon$ is guaranteed.

If we have a particular optimization problem $\mathcal{P}$ that we want to solve up to accuracy ϵ, it could be natural to look for the best method $\mathcal{M}$ for $\mathcal{P}$, i.e. the method having the smallest complexity on this particular problem.

However, this question is not well-posed. Indeed, let us consider the method that outputs, for any problem, the point $\overline{y} \in Q$. When applied to a problem for which $\overline{y}$ is not an optimal solution, this method is completely wrong but in the contrary case (i.e. when $\overline{y} \in X^*$), this method is unbeatable since it solves the problem (exactly!) in a single iteration. For any particular problem $\mathcal{P}$, there exists a trivial optimal method but we are not able to find this method without knowing a priori the optimal solution set.

Furthermore, numerical methods are usually developed for solving many different problems with similar characteristics, not for a particular problem. It is

therefore natural to measure the efficiency of a numerical method $\mathcal{M}$ by the performance of this method on the class of problems for which it has been designed.

Let $\mathbb{P}$ be a class of problems sharing similar characteristics. (We will consider four important classes of convex problems in Section 2.3.)

We can introduce the notion of convergence rate and complexity of a method $\mathcal{M}$ on the class $\mathbb{P}$.

Definition 2.8. Given an optimality measure $Opt(.)$, the convergence rate of the method $\mathcal{M}$ on the class $\mathbb{P}$ is defined as the worst-case convergence rate of the method $\mathcal{M}$ when applied to problems $\mathcal{P}$ from the class $\mathbb{P}$:

$$ConvRate_{\mathbb{P}}(\mathcal{M}, k) = sup_{\mathcal{P}\in\mathbb{P}} ConvRate(\mathcal{P}, \mathcal{M}, k).$$

Definition 2.9. Given an optimality measure $Opt(.)$ and a target accuracy $\epsilon > 0$, the complexity of the method $\mathcal{M}$ on the class $\mathbb{P}$ is the worst-case complexity of the method $\mathcal{M}$ on the problems $\mathcal{P} \in \mathbb{P}$:

$$Compl_{\mathbb{P}}(\mathcal{M}, \epsilon) = \sup_{\mathcal{P}\in\mathbb{P}} Compl(\mathcal{P}, \mathcal{M}, \epsilon)$$

corresponding to the minimal number of iterations after which $\mathcal{M}$ is able to solve any problem from $\mathbb{P}$ with accuracy ϵ.

Now we can look for the optimal method on the class $\mathbb{P}$. Let $\mathbb{M}$ be a family of numerical methods.

Definition 2.10. We define the optimal complexity of the family $\mathbb{M}$ on the class $\mathbb{P}$ as

$$Compl_{\mathbb{M},\mathbb{P}}(\epsilon) = \inf_{\mathcal{M}\in\mathbb{M}} Compl_{\mathbb{P}}(\mathcal{M}, \epsilon).$$

Definition 2.11. A method $\mathcal{M} \in \mathbb{M}$ is an optimal method of the family $\mathbb{M}$ on the class $\mathbb{P}$ if

$$Compl_{\mathbb{P}}(\mathcal{M}, \epsilon) = \Theta(Compl_{\mathbb{M},\mathbb{P}}(\epsilon)), \quad \forall \epsilon \geq 0.$$

The notion of complexity that we have considered for the moment, also known as the analytical complexity, takes only into account the number of iterations that we have to perform to reach a target accuracy, not the individual cost (i.e. the needed number of basic arithmetic operations) of each iteration.

The arithmetical complexity that counts the total number of basic arithmetic operations performed by the numerical method in order to reach a given target accuracy (i.e. the analytical complexity time the cost of each iteration) seems

perhaps a better way to estimate the real efficiency of the method.

However, estimating the cost of each iteration is not an easy task. It depends on the particular structures of f and Q, on the way the oracle and the subproblems used in the scheme are implemented... Furthermore, for methods of the same family (like first-order methods), the cost of each iteration is generally similar. Therefore, when we want to compare different methods from the same family, we can restrict ourselves to the notion of analytical complexity.

We will only use the arithmetical complexity in order to justify the choice of the family of first-order methods (instead of second-order methods for example) for large-scale convex problems.

2.2.3 Large-scale assumption and consequences

In this thesis, we are interested in first-order methods for solving large-scale convex problems. Let us explain what we mean by large-scale and the reasons for the choice of this particular family of methods.

In [65], Nesterov suggests the following subdivision of convex optimization problems depending on the maximum dependence of the iteration cost on the problem size n that we can accept:

1. **Small-size problems**
   A small-size problem is as problem for which the problem dimension n is sufficiently small such that we can consider methods with iteration cost proportional to n^4. In this situation, we can therefore use very sophisticated schemes like the Inscribed Ellipsoid Method (see [38]).

2. **Medium-size problems**
   For medium size problems, we can accept an iteration cost proportional to n^3 and therefore, to compute a matrix inversion or solve a linear system at each iteration. Such kind of problem can therefore be solved using efficient second-order methods like the polynomial interior-point methods (see [54]).

3. **Large-scale problems**
   For large-scale problems, we can at most accept iterations with cost proportional to n^2. The polynomial interior-point methods have attractive analytical complexities of order $\sqrt{\nu}\log(\frac{1}{\epsilon})$ (where ν is the parameter of the self-concordant barrier and ϵ the target accuracy) but each iteration requires to solve a linear system. When the matrix is not sufficiently sparse, such operation has a cost proportional to n^3 which is too costly for such scale of problems.

Whereas we can no longer solve linear systems, matrix-vector multiplications are still allowed. This leads us to consider first-order methods. Indeed, the computation of the gradient, at least for quadratic functions, has a cost similar to the cost of a matrix-vector multiplication. First-order methods (more precisely the gradient-type methods that we will describe in details in Sections 2.4 and 2.5) exhibit worse analytical complexity (the dependence in the target accuracy is no longer logarithmic) but cheaper iteration cost. Therefore, when the target accuracy is not too high, first-order methods are often the methods of choice for solving large-scale problems.

4. **Huge-scale problems**
When the number of variables n becomes huge, even a matrix vector multiplication with cost proportional to n^2 is too costly, and we have to consider methods with cheaper iteration cost of order n or even $\log n$. Examples of such methods are the coordinate descent schemes (see [64, 71]), special subgradient methods with sparse updates (see [65]) or with randomization of the matrix-vector product (see [34]). However, for such methods, the total number of iterations necessary for obtaining an approximate solution of the problem, is typically higher than the corresponding number of iterations of the gradient-type scheme. They must be used only in order to reach relatively poor accuracy for huge-scale problems.

Even if the first-order methods can be slow for obtaining solutions with high accuracy, they share a lot of significant advantages when dealing with large-scale problems:

1. They need a small amount of information at each iteration. The cost of the oracle (computation of the gradient) is more reasonable as compared to the case when we have also to compute the Hessian.

2. They need a reasonable amount of auxiliary computations at each iteration. Providing that the feasible set Q is simple (we will come back to this point in the next subsection), first-order methods are based on cheap auxiliary subproblems solved at each iteration.

3. Their analytical complexity is typically independent on the problem dimension. Whereas the cost of each iteration can increase when the number of variables grows, the needed number of iterations is not modified.

   Remark 2.7. In fact, the analytical complexity depends indirectly on the problem scale via the initial distance to the optimal solution set (i.e. $dist(x_0, X^*)$). Often, it becomes more difficult to guarantee a small initial distance to the optimal solution set when the problem size grows. But this dependence is indirect and in general weak.

We conclude that the first-order methods are very suitable for large scale convex problems when high accuracy is not crucial. As in many large-scale problems, the data is anyway corrupted or known only roughly (by the way, the behavior of first-order methods when used with such inexact information is one of the central subjects of this thesis), it makes more sense to look for solutions with moderate accuracy and therefore to choose a first-order method for solving such kind of problem.

2.2.4 Setup for a first-order method

The theoretical and practical performances of the main first-order methods depends on the a priori choice of a setup for the method.

A setup consists in the choice of

1. a norm $\|.\|$ on E
2. a prox-function $d(x)$ i.e. a differentiable and strongly convex function on Q (with respect to $\|.\|_E$)

Let x_0 be the minimizer of d on Q. By translating and scaling d if necessary, we can always ensure that

$$d(x_0) = 0, \quad d(x) \geq \frac{1}{2}\|x - x_0\|_E^2, \quad \forall x \in Q. \tag{2.8}$$

We define also the corresponding Bregman distance:

$$V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle. \tag{2.9}$$

Due to the strong convexity of $d(x)$ with parameter 1, we have clearly:

$$V(x, z) \geq \frac{1}{2}\|x - z\|_E^2, \quad \forall x, z \in Q. \tag{2.10}$$

All the first-order methods that we will consider are based on subproblems of the form

$$\min_{x \in Q}\{\langle g, x \rangle + \beta d(x)\}$$

with $g \in E^*$, $\beta \in \mathbb{R}_+$. In order to have iterations with moderate cost, the prox-function must be chosen such that this kind of auxiliary subproblem can be solved easily.

On the other hand, the analytical complexity of first-order methods using d as prox-function, depends directly on $d(x^*)$ (where x^* is an arbitrary optimal solution) or on $V(x^*, x_0)$. In order to have a method with good complexity,

the prox-function must be also chosen such that these quantities are as small as possible.

Finding a good setup for which it is easy to minimize functions of the form $\langle g, x\rangle + d(x)$ (with $g \in E^*$) and for which the quantity $d(x^*)$ (or $V(x^*, x_0)$) is reasonably small, is only possible if the feasible set is sufficiently simple.

Typically, constrained problems can be solved by first-order method without modification of the (analytical) complexity compared to the unconstrained case. However, in order to avoid that the auxiliary subproblems become intractable or too costly, we must restrict ourselves to simple sets for which a good setup can be found.

Example 2.1. When $E = \mathbb{R}^n$, two classical setups are:

1. The Euclidean setup: $\|.\|_E = \|.\|_2 = \sqrt{\sum_{i=1}^n (x^{(i)})^2}$ and $d(x) = \frac{1}{2}\|x - x_0\|_2^2$ with $x_0 \in Q$. In this case, the prox-center is x_0 and $V(x, z) = \frac{1}{2}\|x - z\|_2^2$. Furthermore, problems of the form (with $g \in \mathbb{R}^n$)
$$\min_{x \in Q}\{\langle g, x\rangle + \frac{1}{2}\|x - z\|_2^2\} \tag{2.11}$$
can be solved in closed-form with optimal solution:
$$x_{opt} = \pi_Q(z - g)$$
where π_Q denotes the Euclidean projection operator on the set Q.

2. When $Q = \Delta_n = \{x \in \mathbb{R}^n_+, \sum_{i=1}^n x^{(i)} = 1\}$, the l_1 setup (see for example [59]): $\|.\|_E = \|.\|_1 = \sum_{i=1}^n \left|x^{(i)}\right|$ and $d(x) = \ln(n) + \sum_{i=1}^n x^{(i)} \ln(x^{(i)})$ (entropy distance). In this case, the prox-center is $x_0 = \left(\frac{1}{n}, ..., \frac{1}{n}\right)^T$ and $V(x, z) = \sum_{i=1}^n x^{(i)} \ln\left(\frac{x^{(i)}}{z^{(i)}}\right)$. Furthermore, problems of the form (with $g \in \mathbb{R}^n$)
$$\min_{x \in \Delta_n}\{\langle g, x\rangle + d(x)\} \tag{2.12}$$
can be solved in closed-form with optimal solution:
$$x^i_{opt} = \frac{\exp(-g^i)}{\sum_{j=1}^n \exp(-g^j)}, \quad i = 1, ..., n.$$
In the same way, problems of the form (with $g \in \mathbb{R}^n$ and $z \in \Delta_n$)
$$\min_{x \in \Delta_n}\{\langle g, x\rangle + V(x, z)\} \tag{2.13}$$

can be solved in closed-form with optimal solution:

$$x^i_{opt} = \frac{z^i \exp(-g^i)}{\sum_{j=1}^n z^j \exp(-g^j)}, \quad i = 1, ..., n. \tag{2.14}$$

2.3 Classes of Convex Problems

In this thesis, we are interested in first-order methods i.e. methods that solve optimization problems using information coming from a first-order oracle. The classes of convex problems that we consider must be therefore defined by some conditions on the function f that we want to minimize and on its gradient (or its subgradients when the function is non-differentiable).

More precisely, we consider classes of convex problems that are defined using two kinds of assumptions:

1. An assumption on the level of convexity of the function f.
2. An assumption on the level of smoothness of the function f.

2.3.1 Convexity assumption

We consider, mainly, two different levels of convexity:

- f **is convex** on Q. When the function is differentiable, the convexity of f is equivalent to the condition

  $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in Q.$$

  For a nonsmooth function, we have by definition of the subgradients the same kind of inequality

  $$f(y) \geq f(x) + \langle g(x), y - x \rangle, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

  It means that using the answer of a first-order oracle for f at the point x i.e. $\mathcal{O}_1(x) = (f(x), g(x))$, we can construct a linear function $y \to f(x) + \langle g(x), y - x \rangle$ which is a global lower bound for f.

- f is **strongly convex** with parameter $\mu > 0$. When the function is differentiable, the strong convexity of f is equivalent with the condition

  $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_E^2, \quad \forall x, y \in Q.$$

  For a nonsmooth function, the same kind of inequality is satisfied with the gradient replaced by any subgradient

  $$f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_E^2, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

It means that using the answer of a first-order oracle at the point x i.e. $\mathcal{O}_1(x) = (f(x), g(x))$, we can construct a quadratic function: $y \to f(x) + \langle g(x), y - x\rangle + \frac{\mu}{2}\|x - y\|_E^2$ which is a global lower bound for f.

Remark 2.8. The convexity assumptions (f convex or strongly convex) can be seen as a way to ensure the possibility to construct, more or less strong, global lower bound on f using the answer of a first-order oracle.

Remark 2.9. Other levels of convexity can be also considered. The function f is uniformly convex on Q with convexity parameters $\rho \geq 2$ and $\kappa > 0$ if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\kappa}{2}\lambda(1-\lambda)\|x - y\|_E^\rho$$

for all $x, y \in Q$ and for all $\lambda \in [0, 1]$. This condition leads to the following inequality:

$$f(y) \geq f(x) + \langle g(x), y - x\rangle + \frac{\kappa}{2}\|x - y\|_E^\rho, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

When $\rho = 2$, we retrieve the strong convexity. We will come back to the notion of uniform convexity in Chapter 5.

2.3.2 Smoothness assumption

We consider two different levels of smoothness:

- ◇ The **subgradients of f are bounded** i.e. there exists a constant $M < +\infty$ such that

  $$\|g(x)\|_E^* \leq M, \quad \forall x \in Q, \forall g(x) \in \partial f(x).$$

  This condition implies that the subgradients have bounded variations

  $$\|g(x) - g(y)\|_E^* \leq 2M, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x), g(y) \in \partial f(y)$$

  which implies the inequality:

  $$f(y) \leq f(x) + \langle g(x), y - x\rangle + 2M\|x - y\|_E, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

  It means that, using the answer of a first-order oracle at a point $x \in Q$, we can construct a global upper-bound for f of the form linear function + norm. A function with bounded subgradients is not necessarily differentiable. This condition is mainly used in the context of nonsmooth optimization.

Remark 2.10. The boundedness of the subgradients is the most used condition for nonsmooth convex problems (see [74, 68, 55, 58, 63]). Another possible condition is the bounded variation of the subgradients. If a function has bounded subgradients with constant M then these subgradients have bounded variations with constant (at most) $2M$. Indeed $\|g(x) - g(y)\|_E^* \leq \|g(x)\|_E^* + \|g(y)\|_E^* \leq 2M$. On the other hand, when the unconstrained optimization problem $\min_{x \in E} f(x)$ has an optimal solution x^* (and therefore $0 \in \partial f(x^*)$) then the bounded variation of the subgradients with constant M implies the boundedness of the subgradients with the same constant M. Indeed $\|g(x)\|_E^* = \|g(x) - 0\|_E^* \leq M$.

- ◇ The **gradient of f is Lipschitz-continous** i.e. that there exists a constant $L < \infty$ such that

$$\|\nabla f(x) - \nabla f(y)\|_E^* \leq L \|x - y\|_E, \quad \forall x, y \in Q.$$

This condition implies

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

meaning that the the answer of a first-order oracle can be used in order to construct a global quadratic upper-bound for f, of the form linear function + squared norm. A function with Lipschitz-continuous gradient is necessarily differentiable. This condition is classical in the context of smooth optimization.

Remark 2.11. The smoothness assumption can be seen as a way to ensure the possibility to construct global upper-bound on f using the answer of a first-order oracle.

Remark 2.12. We could also consider intermediate levels of smoothness

$$\|g(x) - g(y)\|_E^* \leq L_\nu \|x - y\|_E^\nu .$$

When

- ◇ $\nu = 0$, f has subgradients with bounded variations
- ◇ $\nu = 1$, f has a Lipschitz-continuous gradient
- ◇ $0 < \nu < 1$, f has a Hölder-continuous gradient.

We will come back to these intermediate levels of smoothness later in Chapter 4.

Using the convexity and smoothness assumptions, we are now able to define four classes of optimization problems:

Nonsmooth Convex Problem: $\min_{x \in Q} f(x)$ where

- $\diamond$ Q is convex
- $\diamond$ $f : Q \to \mathbb{R}$ is convex
- $\diamond$ f has bounded subgradients with constant M

Notation: $f \in NC_M(Q)$
Oracle: First-order oracle $\mathcal{O}(x) = (f(x), g(x))$ (where $g(x) \in \partial f(x)$) such that

$$0 \leq f(y) - f(x) - \langle g(x), y - x \rangle \leq 2M \|x - y\|_E, \quad \forall y \in Q.$$

Nonsmooth Strongly Convex Problem: $\min_{x \in Q} f(x)$ where

- $\diamond$ Q is convex
- $\diamond$ $f : Q \to \mathbb{R}$ is strongly convex with parameter $\mu > 0$
- $\diamond$ f has bounded subgradients with constant M

Notation: $f \in NS_{\mu,M}(Q)$
Oracle: First-order oracle $\mathcal{O}(x) = (f(x), g(x))$ (where $g(x) \in \partial f(x)$) such that

$$\frac{\mu}{2} \|x - y\|_E^2 \leq f(y) - f(x) - \langle g(x), y - x \rangle \leq 2M \|x - y\|_E, \quad \forall y \in Q. \tag{2.15}$$

Remark 2.13. In view of inequality (2.15), the class $NS_{\mu,M}(Q)$ is empty when Q is unbounded.

Smooth Convex Problem: $\min_{x \in Q} f(x)$ where

- $\diamond$ Q is convex
- $\diamond$ $f : Q \to \mathbb{R}$ is convex
- $\diamond$ f has a Lipschitz-continuous gradient with constant L

Notation: $f \in SC_L(Q)$
Oracle: First-order oracle $\mathcal{O}(x) = (f(x), \nabla f(x))$ such that

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|x - y\|_E^2, \quad \forall y \in Q.$$

Remark 2.14. In the literature, the class $SC_L(Q)$ is often denoted by $F_L^{1,1}(Q)$ (see [58] for example). For a better consistency with the litterature, we use this second notation in the thesis.

Smooth Strongly Convex Problem: $\min_{x\in Q} f(x)$ where

- $\diamond$ Q is convex
- $\diamond$ $f : Q \to \mathbb{R}$ is strongly convex with parameter $\mu > 0$
- $\diamond$ f has a Lipschitz-continuous gradient with constant L

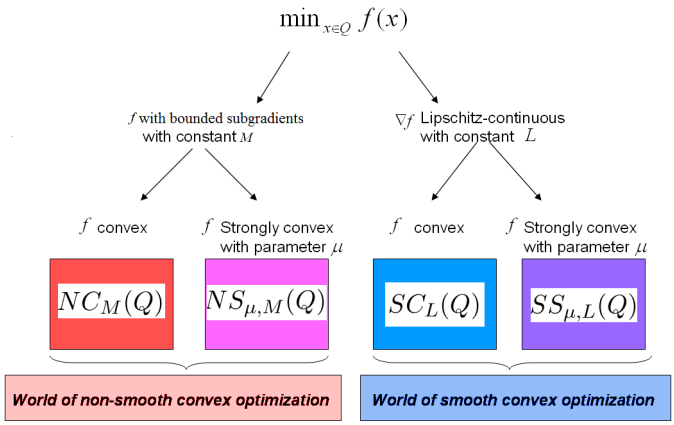
Notation: $f \in SS_{\mu,L}(Q)$
Oracle: First-order oracle $\mathcal{O}(x) = (f(x), \nabla f(x))$ such that

$$\frac{\mu}{2}\|x-y\|_E^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{L}{2}\|x-y\|_E^2, \quad \forall y \in Q. \tag{2.16}$$

Remark 2.15. In the literature, the class $SS_{\mu,L}(Q)$ is often denoted by $S_{\mu,L}^{1,1}(Q)$ (see [58] for example). For a better consistency with the litterature, we use this second notation in the thesis.

Remark 2.16. In view of the inequality (2.16), the class $S_{\mu,L}^{1,1}(Q)$ is empty when $\mu > L$. The Lipschitz-constant of the gradient L is always bigger than or equal to the strong convexity parameter μ. The condition number $Cond := \frac{L}{\mu}$ is always bigger than or equal to one.

As strong convexity is a particular case of convexity, smooth convex problems and smooth strongly convex problems form together the world of smooth convex optimization. Similarly, nonsmooth convex and nonsmooth strongly convex problems form the world on nonsmooth convex optimization.



2.4 First-Order Methods in Smooth Convex Optimization

In this section, we are interested in first-order methods in the context of smooth convex optimization i.e. numerical methods developed in order to minimize a smooth convex (or smooth strongly convex) function using a first-order oracle. The goal of such methods is to generate a sequence of approximate solutions $\{y_k\}_{k\geq 0}$, constructed by using only first-order information $(f(x_k), \nabla f(x_k))$ at some search points $\{x_k\}_{k\geq 0}$, such that the accuracy $Opt(y_k)$ converges to zero as quickly as possible.

Before introducing three important first-order methods of smooth convex optimization, let us study what kind of performance can be expected from such methods.

2.4.1 Optimal complexity

Convex case

Let us start with smooth convex problems. For a fixed $L < \infty$, we consider $\mathbb{P}_{SC}(L)$, the class of optimization problems of the form $\min_{x\in Q} f(x)$ where Q is any convex set in E, E is any finite-dimensional vector space and f is any function in $F_L^{1,1}(Q)$ endowed with an exact first-order oracle.

For solving such class of problems, we consider $\mathbb{M}_{SC}(L, R)$ the family of black-box first-order methods applicable to any problem in $\mathbb{P}_{SC}(L)$, such that when applied to a problem $\mathcal{P} \in \mathbb{P}_{SC}(L)$, it starts with an initial point x_0 satisfying $dist(x_0, X^*) \leq R$ (where X^* is the optimal solution set of problem $\mathcal{P}$).

Remark 2.17. When the Euclidean setup is used, we have $d(x^*) = \frac{1}{2}\|x_0 - x^*\|_E^2 \leq \frac{1}{2}R^2$ (where x^* is the point in the optimal solution set closest to x_0). In general, we have only the inequality $d(x^*) \geq \frac{1}{2}\|x_0 - x^*\|^2$. By abusing the notation, for the analysis of our first-order methods, we denote by $\frac{1}{2}R^2$ the quantity $d(x^*)$ that represents in some sense the squared distance between the initial point x_0 (which is the minimizer of the prox-function) and the optimal solution x^*. As $d(x_0) = 0$ and $\langle \nabla d(x_0), x^* - x_0\rangle \geq 0$, we have:

$$V(x^*, x_0) \leq d(x^*) = \frac{1}{2}R^2.$$

Remark 2.18. When the feasible set Q is bounded and $d(x^*) = \frac{1}{2}\|x_0 - x^*\|_E^2 = \frac{1}{2}R^2$, the initial distance to the optimal solution set can be of course bounded by the diameter D of Q, such that we have always $R \leq D$.

The behavior of a black-box first-order method in $\mathbb{M}_{SC}(L,R)$ cannot be arbitrarily good. The best performance that we can expect is given by the following lower bound on the optimal complexity of first-order methods for smooth convex problems:

Theorem 2.8. *([58, 55]) Using the optimality measure $Opt_f(y) = f(y) - f^*$, the complexity $Compl_{\mathbb{P}_{SC}(L)}(\mathcal{M},\epsilon)$, of a first-order method $\mathcal{M}$ in $\mathbb{M}_{SC}(L,R)$ on the problem class $\mathbb{P}_{SC}(L)$ cannot be better than*

$$O(1)\sqrt{\frac{LR^2}{\epsilon}}$$

where $O(1)$ denotes an absolute constant factor independent of L, of R and of ϵ.

Remark 2.19. Since there is no assumption on the dimension of the problem in the definition of our problem class, any complexity bound must be dimension-independent. When the dimension of E is fixed, it is possible to obtain lower complexity bounds with a better dependence in ϵ but with a direct dependence on n, the dimension of E. As we are interested in solving large-scale problems (problems for which it is impossible to perform a number of iterations of the same order as the problem size), we restrict ourselves by numerical methods having an analytical complexity independent of the problem dimension.

As a consequence of the previous Theorem

Theorem 2.9. *[58, 55] The convergence rate $ConvRate_{\mathbb{P}_{SC}(L)}(\mathcal{M},\epsilon)$, of a first-order method $\mathcal{M} \in \mathbb{M}_{SC}(L,R)$, on the problem class $\mathbb{P}_{SC}(L)$, cannot be better than*

$$O(1)\frac{LR^2}{k^2}$$

where $O(1)$ denotes an absolute constant factor independent of L, of R and of k.

We conclude that the optimal complexity of $\mathbb{M}_{SC}(L,R)$ on $\mathbb{P}_{SC}(L)$ is such that

$$Compl_{\mathbb{M}_{SC}(L,R),\mathbb{P}_{SC}(L)}(\epsilon) \geq O(1)\sqrt{\frac{LR^2}{\epsilon}}.$$

We will see later that this inequality is in fact an equality: it is possible to develop a first-order method (the fast gradient method, see subsection 2.4.5) belonging to $\mathbb{M}_{SC}(L,R)$ and that exhibits exactly such optimal complexity $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$.

Strongly convex case

For fixed $0 < \mu \leq L < +\infty$, we consider $\mathbb{P}_{SS}(L, \mu)$, the class of optimization problems of the form $\min_{x \in Q} f(x)$ where Q is any convex set in E, E is any finite-dimensional vector space and f is any function in $S^{1,1}_{\mu,L}(Q)$ endowed with an exact first-order oracle.

For solving such class of problems, we consider $\mathbb{M}_{SS}(L, \mu, R)$ the family of black-box first-order method applicable to any problem in $\mathbb{P}_{SS}(L, \mu)$, such that when applied to a problem $\mathcal{P} \in \mathbb{P}_{SS}(L, \mu)$, it starts with an initial point x_0 satisfying $\|x_0 - x^*\|_E \leq R$ (with x^* the unique optimal solution of problem $\mathcal{P}$).

When the function is strongly convex, we can expect from a first-order method, a significantly better behavior than what is possible when $\mu = 0$:

Theorem 2.10. *([58, 55]) Using the optimality measure $Opt_f(y) = f(y) - f^*$, the complexity $Compl_{\mathbb{P}_{SS}(L,\mu)}(\mathcal{M}, \epsilon)$, of a first-order method $\mathcal{M} \in \mathbb{M}_{SS}(L, \mu, R)$, on the problem class $\mathbb{P}_{SS}(L, \mu)$, cannot be better than*

$$O(1)\sqrt{\frac{L}{\mu}} \log\left(\frac{\mu R^2}{\epsilon}\right)$$

where $O(1)$ denotes an absolute constant factor independent of L, of μ, of R and of ϵ.

This lower-bound on the optimal complexity depends on ϵ only logarithmically. For a method exhibiting such complexity, the target accuracy can be chosen very small without big effect on the needed number of iterations. We will see that such method exists (see subsection 2.4.5), implying that

$$Compl_{\mathbb{M}_{SS}(L,\mu,R),\mathbb{P}_{SS}(L,\mu)}(\epsilon) = O(1)\sqrt{\frac{L}{\mu}} \log\left(\frac{\mu R^2}{\epsilon}\right).$$

We conclude that for the complexity of a first-order method on smooth strongly convex problems, the central quantity is not the target accuracy ϵ but the condition number of the problem $Cond := \frac{L}{\mu}$.

2.4.2 Primal Gradient Method (PGM)

Convex Case

The (primal) gradient method is certainly the most natural and the most well-known, first-order method for smooth problems. For the moment, let us work with the Euclidean setup. In the unconstrained case (i.e. when Q is the whole

space E), the principle of the gradient method is to go, from the current search point x_k, along the direction of the anti-gradient, the direction of steepest-descent:

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

where $\{\gamma_k\}_{k\geq 0} \subset \mathbb{R}^+$ is a sequence of stepsizes. Various stepsizes strategies can be used in the PGM: constant or decreasing sequence chosen in advance, line search with a full relaxation or using for example the Goldstein-Armijo rule (see for example section 1.2.3 in [58] and chapter 3 in [66]).

In the context of smooth convex optimization, the behavior of the function is much more predictable that in the nonconvex or nonsmooth cases. In term of worst case behavior, we can restrict ourselves w.l.o.g. to constant stepsize (i.e. $\gamma_k = \gamma$ for all $k \geq 0$). When $f \in F_L^{1,1}(Q)$, the optimal stepsize choice (i.e. the choice minimizing the complexity of the method) is given by $\gamma = \frac{1}{L}$ (see section 2.1.5 in [58]), for which the gradient step becomes

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k). \tag{2.17}$$

Remark 2.20. When the objective function belongs to $F_L^{1,1}(Q)$ but the Lipschitz-constant of the gradient is not known, the Primal Gradient Method can be used with an initial optimistic estimate of the Lipschitz constant $L_0 \leq L$ coupled with a backtracking procedure (see Section 3 in [62]).

The gradient method can be easily generalized to the constrained case. Indeed, in the unconstrained case, the gradient step (2.17) is nothing else but the minimizer on E of the quadratic function

$$Q_{L,x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_E^2$$

(which is a global upper bound on f since $f \in F_L^{1,1}(E)$).

Therefore, it is natural to generalize the gradient step in the constrained case by

$$x_{k+1} = T_L(x_k, \nabla f(x_k)) \tag{2.18}$$

where

$$T_L(z, g) = \arg\min_{x \in Q} \{f(z) + \langle g, x - z \rangle + \frac{L}{2} \|x - z\|_E^2\} \tag{2.19}$$

for any $z \in E$ and $g \in E^*$.

Furthermore, working with the Euclidean norm, simple algebra shows that

$$Q_{L,x_k}(x) = \frac{L}{2} \left\| x - (x_k - \frac{1}{L}\nabla f(x_k)) \right\|_2^2 - \frac{1}{2L} \|\nabla f(x_k)\|_2^2 + f(x_k)$$

which implies that the minimizer of this function on Q can be also obtained as the orthogonal projection of the gradient step on Q

$$x_{k+1} = \pi_Q(x_k - \frac{1}{L}\nabla f(x_k))$$

where π_Q denotes the orthogonal projection operator on Q.

We are now able to give the complete scheme of the (Euclidean) primal gradient method:

Algorithm 2 Primal Gradient Method (PGM): Euclidean version

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $\mathcal{O}(x_k) = (f(x_k), \nabla f(x_k))$.
4: Compute $x_{k+1} = T_L(x_k, \nabla f(x_k)) = \pi_Q(x_k - \frac{1}{L}\nabla f(x_k))$.
5: **end for**

During the execution of the scheme, we generate only a sequence of search points $\{x_k\}_{k\geq 0}$ that can be updated independently of any other sequence. But if we want to establish a convergence rate for this method, we have also to define a sequence of approximate solutions $\{y_k\}_{k\geq 0}$ for which a converging behavior can be proved for $Opt_f(y_k) = f(y_k) - f^*$.

An important property of the gradient method is that, if the first-order oracle is exact, the sequence $Opt_f(x_k)$ is decreasing to zero. We can therefore choose for the sequence of approximate solutions $y_k = x_k$. Another choice, more robust when the first-order information given by the oracle is affected by some noise (see Chapter 4 and 7), is to choose an averaging of the search points

$$y_k = \frac{\sum_{i=1}^k x_i}{k}.$$

It is well-known (see for example [58, 61]) that both choices lead to a convergence rate of the form $\Theta\left(\frac{LR^2}{k}\right)$. More precisely, we have

Theorem 2.11. *Assume that $f \in F_L^{1,1}(Q)$ is endowed with an exact first-order oracle. Then the sequence $y_k = \frac{\sum_{i=1}^k x_i}{k}$ (or $y_k = \arg\min_{x_1,\ldots,x_k} f(x_i) = x_k$) generated by the Primal Gradient Method satisfies*

$$f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

Proof. See for example Theorem 4.3 of Chapter 4 with $\delta = 0$. □

In view of this convergence rate, obtaining an ϵ-solution with the Primal Gradient Method (PGM) takes $O\left(\frac{LR^2}{\epsilon}\right)$ iterations. Comparing this complexity with the optimal complexity given by Theorem 2.8, we conclude that the PGM is not an optimal first-order method for smooth convex problems. For such kind of problems, it is possible to develop more efficient first-order methods as we will see in subsection 2.4.5.

In some cases, the Lipschitz-constant of ∇f can be smaller when working with another norm than the Euclidean one. Furthermore, the set Q can be such that it is difficult to minimize quadratic functions of the form $Q_{L,x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_E^2$ over this set. Therefore, it could be interesting to generalize the PGM to non-Euclidean setups, keeping the same complexity and adding a degree of freedom for the choice of the norm and the choice of the prox-function.

Let us choose a norm $\|.\|_E$ on E and a prox-function $d(.)$. Inspired by non-Euclidean subgradient methods in nonsmooth convex optimization (see [3, 36] and section 2.5.2), we propose here a generalization of PGM to non-Euclidean setup:

Algorithm 3 Primal Gradient Method (PGM): Non-Euclidean version

1: Compute $x_0 =$ arg $\min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $\mathcal{O}(x_k) = (f(x_k), \nabla f(x_k))$.
4: Compute $x_{k+1} =$ arg $\min_{x \in Q} [\langle \nabla f(x_k), x - x_k \rangle + LV(x, x_k)]$
5: **end for**

The convergence rate of this non-Euclidean gradient method is also of the form $O\left(\frac{LR^2}{k}\right)$:

Theorem 2.12. *Assume that $f \in F_L^{1,1}(Q)$ is endowed with an exact first-order oracle. Then the sequence $y_k = \frac{\sum_{i=1}^{k} x_i}{k}$ generated by the non-Euclidean Primal Gradient Method satisfies*

$$f(y_k) - f^* \leq \frac{LV(x^*, x_0)}{k}$$

where x^* is any optimal solution or problem $\mathcal{P}$.

Proof. See the proofs of Theorems 7.1 and 7.2 in Chapter 7 with $\phi = f$, $\delta = 0$, $\sigma = 0$ and $\gamma_i = \frac{1}{L}$ for all i. □

When the Euclidean norm is used with the prox-function $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, we have $V(x,z) = \frac{1}{2}\|x - z\|_2^2$. We retrieve the Euclidean PGM and the convergence rate given by Theorem 2.11 .

In view of the non-Euclidean PGM and of its convergence rate given by Theorem 2.12, we conclude that the setup must be chosen such that

- $\diamond$ the Lipschitz-constant of ∇f with respect to the norm $\|.\|_E$ is small
- $\diamond$ the value of $V(x^*, x_0)$ is small where x^* is an arbitrary optimal solution of problem $\mathcal{P}$
- $\diamond$ subproblems of the form $\arg\min_{x\in Q}[\langle g, x - z\rangle + V(x,z)]$ are easy to solve for any $z \in E$ and any $g \in E^*$.

Remark 2.21. To the best of our knowledge, it is the first time that the Primal Gradient Method is explicitly generalized to the non-Euclidean case. The reader can find after the conclusion, a table of methods, emphasizing the new schemes that have been developed in this thesis.

Strongly Convex Case

Let us now look at the smooth strongly convex case i.e. when the function to be minimized belongs to $S^{1,1}_{\mu,L}(Q)$ (with $0 < \mu \leq L < +\infty$). In this strongly convex case, we restrict ourselves to the Euclidean setup.

One important property of the PGM is that the strong convexity parameter μ does not need to be used explicitly in the scheme. The Primal Gradient Method for strongly convex problems is exactly the same scheme as in the convex case.

However, even if the scheme is the same, strong convexity can lead to a significant acceleration of the method. Let us choose for the approximate solution $y_k = \arg\min_{x_1,\ldots,x_k} f(x_k)$ ($= x_k$ since the method is monotone when the oracle is exact). The convergence of the PGM in the smooth strongly convex case is given by

Theorem 2.13. *Assume that $f \in S^{1,1}_{\mu,L}(Q)$ is endowed with an exact first-order oracle then the sequence $\{y_k\}_{k\geq 0} = \arg\min_{x_1,\ldots,x_k} f(x_k)$ generated by the Primal Gradient Method satisfies*

$$f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2}\exp\left(-k\frac{\mu}{L}\right).$$

Proof. See for example Theorem 5.4 of Chapter 5 with $\delta = 0$. □

We see that the optimality measure decreases exponentially with the iteration counter k, leading to a much faster convergence rate than in the non-strongly convex case (at least when μ is not too small). In term of complexity, an ϵ-solution can be obtained after

$$O\left(\frac{L}{\mu}\ln\left(\frac{LR^2}{\epsilon}\right)\right) \tag{2.20}$$

iterations. Even when μ is not known, the strong convexity can be exploited and we obtain a complexity which is significantly better (at least when $\frac{\mu}{L}$ is not too small) than $O\left(\frac{LR^2}{\epsilon}\right)$, obtained in the smooth non-strongly convex case.

However, the complexity of the PGM in the strongly convex case $O\left(\frac{L}{\mu}\ln\left(\frac{LR^2}{\epsilon}\right)\right)$ depends linearly on the condition number $\frac{L}{\mu}$. This dependence is not optimal in view of Theorem 2.10. The optimal complexity of first-order methods on smooth strongly convex problems, i.e. $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\mu R^2}{\epsilon}\right)\right)$, depends on the square root of the condition number, not on the condition number itself. We conclude that the PGM is not an optimal first-order method for smooth strongly convex problems: it is possible to do better, as we will see with the fast gradient method in subsection 2.4.5.

Remark 2.22. When the strong convexity parameter μ and the number of iterations k are small, the convergence rate $f(y_k) - f^* \leq \frac{LR^2}{2}\exp\left(-k\frac{\mu}{L}\right)$ obtained in Theorem 2.13, assuming that $f \in S^{1,1}_{\mu,L}(Q)$, can be worse that the convergence rate $f(y_k) - f^* \leq \frac{LR^2}{2k}$ obtained in Theorem 2.11, assuming that $f \in F^{1,1}_{L}(Q)$. However, as the scheme is exactly the same in both cases and as $S^{1,1}_{\mu,L}(Q) \subset F^{1,1}_{L}(Q)$, the second bound is also valid in the strongly convex case and we obtain the convergence rate

$$f(y_k) - f^* \leq \frac{LR^2}{2}\min\left(\exp\left(-k\frac{\mu}{L}\right), \frac{1}{k}\right).$$

2.4.3 The machinery of estimate functions

The most recent and efficient first-order methods in smooth convex optimization are based on the machinery of estimate functions (see [58, 59, 62]).

The principle of this approach is to construct progressively, using two sequences of coefficients $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$,

1. a model $\Psi_k(x)$ of the function using typically all the previously accumulated first-order information,

2. a sequence of approximate solutions y_k (for which we obtain the convergence rate)

such that the two following inequalities are satisfied:

$$A_k f(y_k) \leq \Psi_k^* = \min_{x \in Q} \Psi_k(x) \text{ and } \Psi_k(x) \leq A_k f(x) + \beta_k d(x), \quad \forall x \in Q$$

where $A_k = \sum_{i=0}^{k} \alpha_i$ and $d(x)$ is the prox-function chosen in the setup.
The convergence rate depends directly on the two sequences of coefficients. Indeed, $A_k f(y_k) \leq \Psi_k^* \leq \Psi_k(x^*) \leq A_k f^* + \beta_k d(x^*)$, and therefore: $f(y_k) - f^* \leq \frac{\beta_k d(x^*)}{A_k}$.

Remark 2.23. The fact that the model $\Psi_k(x)$ is based on all previously accumulated first-order information during the k first steps of the scheme does not mean that we have to store all these data in memory (like what is needed for classical bundle methods in nonsmooth optimization). Typically, we only have to store and update a weighted sum of the accumulated gradients.

The methods based on this principle typically update different sequences of iterates:

- ◇ a sequence of search points x_k, where we compute the first-order information,
- ◇ the sequence $z_k = \arg\min_{x \in Q} \Psi_k(x)$ of minimizers of the estimate functions $\Psi_k(x)$,
- ◇ a sequence of approximate solutions y_k, for which we obtain the convergence rate,
- ◇ sometimes, one or more additional sequences, often obtained using gradient steps.

The easiest way to implement the idea of sequences of estimate functions is the Dual Gradient Method (DGM) introduced by Nesterov in [62]. This method shares the same behavior as the Primal Gradient Method (PGM). A more sophisticated implementation of this machinery leads to the Fast Gradient Method (FGM), developed by Nesterov in different versions [58, 59, 62] and that can reach the optimal convergence rate for smooth convex problems and smooth strongly convex problems.

Remark 2.24. In the deterministic case (i.e. when the information given by the first-order oracle is deterministic), β_k is typically chosen equal to L, at least when this Lipschitz-constant of the gradient is known. When the objective function belongs to $F_L^{1,1}(Q)$ but the Lipschitz-constant of the gradient is not

known, the Dual Gradient Method and the Fast Gradient Method can be used with an initial optimistic estimate of the Lipschitz constant $L_0 \leq L$ coupled with a backtracking procedure (see Section 4 in [62] and [72]).

Remark 2.25. When the first-order oracle is stochastic, we will see in Chapter 7 that this constant coefficient policy $\beta_i = L$ is not the best choice anymore.

Let us now look at the Dual Gradient Method and Fast Gradient Method in details.

2.4.4 Dual Gradient Method (DGM)

Convex Case

The Dual Gradient Method (DGM) is the easiest way to implement the idea of estimate functions. This method has been introduced in [62] for smooth convex problems, using a Euclidean setup. In the DGM, the model, updated at each iteration, has the form

$$\Psi_k(x) = \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{2} \|x - x_0\|_2^2 \tag{2.21}$$

where $\{x_i\}_{i \geq 0}$ is the sequence of search points generated by the method and $\{\alpha_i\}_{i \geq 0}$ is a sequence of coefficients satisfying $\alpha_i \leq 1$ for all $i \geq 0$.

The new search point x_{k+1} is simply chosen as the minimizer of the model

$$x_{k+1} = \arg\min_{x \in Q} \Psi_k(x).$$

The sequence $\{x_k\}_{k \geq 0}$ can therefore be updated without the help of any other sequence of iterates (a property shared also by the PGM). However, in order to define a sequence of approximate solutions for which an upper-bound on $Opt_f(y_k)$ can be guaranteed, additional work must be done. From each search point x_i $(0 \leq i \leq k)$, separately, let us take a gradient step $w_i = T_L(x_i, \nabla f(x_i))$. Averaging the obtained points, we can define the approximate solution

$$y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{A_k}$$

where $A_k = \sum_{i=0}^{k} \alpha_i$.

Remark 2.26. At each iteration i, we must compute the search point x_i and the additional point w_i. The approximate solution y_k must be only computed when we stop the scheme.

We are now able to describe the complete scheme of the Dual Gradient Method:

Algorithm 4 Dual Gradient Method (DGM): Euclidean version

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f(x_k), \nabla f(x_k))$.
4: Compute $w_k = T_L(x_k, \nabla f(x_k))$
5: Compute $x_{k+1} = \arg\min_{x \in Q} \left[\sum_{i=0}^{k} \alpha_i [\langle \nabla f(x_i), x - x_i \rangle] + \frac{L}{2} \|x - x_0\|_2^2 \right]$.
6: **end for**

The sequences $\{\Psi_k(x)\}_{k \geq 0}$ and $\{y_k\}_{k \geq 0}$ define a sequence of estimate functions (see [62]) and the DGM exhibits therefore the convergence rate

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k}$$

where $A_k = \sum_{i=0}^{k} \alpha_i$ and $d(x^*) = \frac{\|x_0 - x^*\|_2^2}{2}$(since the Euclidean setup is used).

In view of this result and of the condition $\alpha_i \leq 1$, we conclude that the optimal coefficients choice is given by $\alpha_i = 1$ for all $i \geq 0$. With this choice, we obtain

Theorem 2.14. *Assume that $f \in F_L^{1,1}(Q)$ is endowed with an exact first-order oracle then the sequence $y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{A_k}$, generated using the Dual Gradient Method with $\alpha_i = 1$ for all $i \geq 0$, satisfies*

$$f(y_k) - f^* \leq \frac{L \|x_0 - x^*\|_2^2}{2(k+1)}.$$

This convergence rate is completely similar to what we have obtained for the Primal Gradient Method. Like the PGM, the DGM exhibits the same non-optimal complexity $O\left(\frac{LR^2}{\epsilon}\right)$ on smooth convex problems.

Similarly to what we have done for the PGM, we propose now, for the first time in the literature, a generalization of the Dual Gradient Method to non-Euclidean setups. The idea is to replace the squared norm term in the model by a prox-function

$$\Psi_k(x) = \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] + Ld(x) \tag{2.22}$$

and the gradient step defining w_k by the generalization already used for the PGM

$$w_k = \arg\min_{x \in Q} \{LV(x, x_k) + \langle \nabla f(x_k), x - x_k \rangle\}.$$

In order to initialize the recurrence $A_0 f(y_0) \leq \Psi_0^* = \min_{x \in Q} \Psi_0(x)$, w_0 must be defined in a slightly different way:

$$w_0 = \arg\min_{x \in Q} \Psi_0(x).$$

Our generalization of the Dual Gradient Method to non-Euclidean setup is therefore given by

Algorithm 5 Dual Gradient Method (DGM): Non-Euclidean version

1: Compute $x_0 = \arg\min_{x \in Q} d(x)$
2: Obtain $(f(x_0), \nabla f(x_0))$
3: Compute $w_0 = \arg\min_{x \in Q} \{Ld(x) + \alpha_0[\langle \nabla f(x_0), x - x_0 \rangle]\}$
4: **for** k = 1 : ... **do**
5: Compute $x_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k-1} \alpha_i[\langle \nabla f(x_i), x - x_i \rangle]\}$
6: Obtain $(f(x_k), \nabla f(x_k))$
7: Compute $w_k = \arg\min_{x \in Q} \{LV(x, x_k) + \langle \nabla f(x_k), x - x_k \rangle\}$.
8: **end for**

Here also the sequences $\{\Psi_k(x)\}_{k \geq 0}$ and $\{y_k\}_{k \geq 0}$ define a sequence of estimate functions (see the proof of Lemma 7.3 in Chapter 7 with $\delta = 0$, $\sigma = 0$, $\phi = f$ and $\beta_i = L$ for all $i \geq 0$) and therefore

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k}.$$

Here also the optimal choice for the coefficients is $\alpha_i = 1$ for all i and we obtain the convergence rate

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{k+1}.$$

When the Euclidean setup is used, we retrieve the Euclidean DGM with its convergence rate given by Theorem 2.14.

Strongly Convex Case

Contrarily to the PGM, the DGM must be adapted in order to take advantage of the strong convexity of f. We propose here, for the first time in the literature, such modification of the DGM to the strongly convex case (restricting ourselves to the Euclidean setup). We use the strong convexity parameter μ

1. In the choice of the coefficients $\{\alpha_i\}_{i \geq 0}$. The sequence is chosen such that

$$\alpha_0 = \frac{L}{L - \mu}, \quad (L - \mu)\alpha_{k+1} = A_k(\mu)\mu + L \tag{2.23}$$

where $A_k(\mu) = \sum_{i=0}^{k} \alpha_i$. When $\mu = 0$, we retrieve the choice $\alpha_i = 1$ for all $i \geq 0$.

2. In the choice of the model of the function:

$$\Psi_k(x) = \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_E^2] + \frac{L}{2}\|x - x_0\|_E^2.$$

When $\mu = 0$, we retrieve the choice used in the previous subsection.

With these modifications, we obtain the following Dual Gradient Method for smooth strongly convex problems

Algorithm 6 Dual Gradient Method (DGM) for strongly convex problems

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f(x_k), \nabla f(x_k))$.
4: Compute $w_k = T_L(x_k, \nabla f(x_k)$
5: Compute

$$x_{k+1} = \arg\min_{x \in Q} \left[\sum_{i=0}^{k} \alpha_i [\langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_E^2] + \frac{L}{2}\|x - x_0\|_E^2 \right]$$

.
6: **end for**

With $y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{A_k}$ or $y_k = \arg\min_{x_0,\ldots,x_k} f(x_i)$, it is possible to prove (see Theorem 5.5 with $\delta = 0$) that the sequences $\{\Psi_k(x)\}_{k\geq 0}$ and $\{y_k\}_{k\geq 0}$ define a sequence of estimate functions. The convergence of the method is therefore given by

$$f(y_k) - f^* \leq \frac{L\,\|x_0 - x^*\|_2^2}{2A_k(\mu)}.$$

This rate seems at first sight to be exactly the same than the one in the non-strongly convex case but the strong convexity allows the coefficients to grow much more quickly with the iteration counter. Indeed, we have $\{A_k(\mu)\}_{k\geq 0} = \sum_{i=1}^{k+1} \left(\frac{L}{L-\mu}\right)^i$ (see Lemma 5.6 in Chapter 5) and therefore

- $A_k = A_k(0) = k + 1$, $\forall k \geq 0$ if $\mu = 0$
- $A_k = A_k(\mu) \geq \left(\frac{L}{L-\mu}\right)^{k+1}$, $\forall k \geq 0$ if $\mu > 0$.

We obtain finally the following theorem

Theorem 2.15. *Assume that $f \in S^{1,1}_{\mu,L}(Q)$ is endowed with an exact first-order oracle. Then the sequence $y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{A_k}$ (or $y_k = \arg\min_{x_0,\dots,x_k} f(x_i)$) generated using the Dual Gradient Method satisfies*

$$f(y_k) - f^* \leq \frac{L \|x_0 - x^*\|_2^2}{2(k+1)}, \text{ if } \mu = 0$$

and

$$f(y_k) - f^* \leq \frac{L \|x_0 - x^*\|_2^2}{2} \left(1 - \frac{\mu}{L}\right)^{k+1} \leq \frac{L \|x_0 - x^*\|_2^2}{2} \exp\left(-(k+1)\frac{\mu}{L}\right)$$

if $\mu > 0$.

We conclude that our generalization of the Dual Gradient Method in the smooth strongly convex case exhibits the same non-optimal complexity $O\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)$ as the PGM.

Remark 2.27. We have $A_k(\mu) \geq A_k(0)$ for all $k \geq 1$ (see remark 5.9 on Chapter 5). Therefore the upper-bound $f(\hat{y}_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2(k+1)}$ is also available in the case $\mu > 0$ and we have

$$f(\hat{y}_k) - f^* \leq \frac{L \|x_0 - x^*\|_2^2}{2} \min\left(\frac{1}{k+1}, \exp\left(-(k+1)\frac{\mu}{L}\right)\right).$$

2.4.5 Fast Gradient Method (FGM)

The methods that we have studied for the moment exhibit a complexity $O\left(\frac{LR^2}{\epsilon}\right)$ for smooth convex problems and $O\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)$ for smooth strongly convex problems. In both cases, such complexities are not optimal.

We are interested now in first-order methods that are able to reach the optimal complexities $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ in the smooth convex case and $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)$ in the smooth strongly convex case. These methods have been developed since 1983 by Nesterov under various variants (see [56, 57, 58, 59, 62]) and are known under the generic name of fast gradient methods (the appellation optimal first-order methods or Nesterov optimal methods are also used).

In this thesis, we consider the version of the Fast Gradient Method introduced by Nesterov in [59]. This method is based, like the Dual Gradient Method, on the machinery of estimate functions but uses it in a more sophisticated and smarter way. Let us start with the smooth convex case.

Convex Case

The Fast Gradient Method introduced in [59] is designed for solving smooth convex problems using an arbitrary setup. Let us therefore choose a norm on E, $\|.\|_E$ and a prox-function $d(.)$.

The method is based on a sequence of positive coefficients $\{\alpha_k\}_{k\geq 0}$ that must satisfy

$$\alpha_0 \quad \in \quad (0,1], \quad \alpha_k^2 \;\leq\; A_k \;\stackrel{\text{def}}{=}\; \sum_{i=0}^{k} \alpha_i, \quad k \geq 0. \tag{2.24}$$

Let us define also $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$, $k \geq 0$.

Like in the DGM, the model of the function is constructed as

$$\Psi_k(x) = Ld(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \tag{2.25}$$

where $\{x_i\}_{i\geq 0}$ is the sequence of search points generated by the method.

However, contrarily to the DGM, the new search point is not defined as the minimizer of the model but as a convex combination

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$$

between

1. The minimizer of the model

$$z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle]\}$$

2. A gradient step taken from x_k: $y_k = T_L(x_k, \nabla f(x_k))$.

This method can be seen as a smart mix between the PGM (where $x_{k+1} = y_k$) and the DGM (where $x_{k+1} = z_k$).

We are now able to describe in details the fast gradient method:

Algorithm 7 Fast Gradient Method (FGM)

1: Compute $x_0 = \min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f(x_k), \nabla f(x_k))$
4: Compute $y_k = T_L(x_k, \nabla f(x_k))$
5: Compute $z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [\langle \nabla f(x_i), x - x_i \rangle]\}$
6: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$ where $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$
7: **end for**

Since the sequence $\{\Psi_k(x)\}_{k \geq 0}$ together with $\{y_k\}_{k \geq 0}$ define a sequence of estimate functions (see [59]), we obtain the convergence rate $f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k}$. At first sight, this seems completely similar to what we have obtained for the DGM. However the condition (2.24) allows us to consider growing coefficients $\{\alpha_i\}_{i \geq 0}$. A simple choice for the sequence $\{\alpha_i\}_{i \geq 0}$ consists in letting $\alpha_i = \frac{i+1}{2}$ for which we have $A_k = \frac{(k+1)(k+2)}{4}$, $\tau_k = \frac{2}{k+3}$, and therefore

Theorem 2.16. *[59] Assume that $f \in F_L^{1,1}(Q)$ is endowed with an exact first-order oracle. Then the sequence y_k generated by the Fast Gradient Method with $\alpha_i = \frac{i+1}{2}$ for all $i \geq 0$, satisfies*

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)^2}.$$

With coefficients growing proportionally to the iteration counter, the fast gradient method is able to reach the optimal convergence rate $O\left(\frac{LR^2}{k^2}\right)$ and therefore the corresponding optimal complexity $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ on smooth convex problems. The FGM is therefore an optimal first-order method in smooth convex optimization.

This method can be used with any norm $\|.\|_E$ and any prox-function d (which must be chosen such that $d(x^*)$ is small and subproblems based on d are easy). However, in this scheme, the subproblem $\min_{x \in Q} \{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_E^2\}$ defining y_k is not based on the prox-function but on the squared norm. Such kind of subproblems can be difficult to solve. In view of this potential difficulty, Nesterov also proposes in [59] a variant of the fast gradient method which use only subproblems based on the prox-function or the corresponding Bregman distance. The scheme looks as follow:

Algorithm 8 Fast Gradient Method (FGM) using Bregman distance

Compute $x_0 = \arg\min_{x \in Q} d(x)$
Obtain $(f(x_0), \nabla f(x_0))$
Compute $y_0 =$ $\arg\min_{x \in Q} \{Ld(x) + \alpha_0 \langle \nabla f(x_0), x - x_0 \rangle\}$
for k = 0 : ... **do**
  Compute $z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i \langle \nabla f(x_i), x - x_i \rangle\}$
  Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$
  Obtain $(f(x_{k+1}), \nabla f(x_{k+1}))$
  Compute $\hat{x}_{k+1} = \arg\min_{x \in Q} \{LV(x, z_k) + \alpha_{k+1} \langle \nabla f(x_{k+1}), x - z_k \rangle\}$
  Compute $y_{k+1} = \tau_k \hat{x}_{k+1} + (1 - \tau_k) y_k$.
end for

This method seems at first sight more complicated but uses only subproblems based on the prox-function $d(x)$ (or equivalently on the corresponding Bregman distance $V(x, z)$). This property can be crucial in some situations, reducing significantly the cost of each iteration. Furthermore, this modification does not change at all the convergence rate, the Theorem 2.16 remains valid.

Remark 2.28. As an example, let us consider the situation where $Q = \{x \in \mathbb{R}^n_+ | \sum_{i=1}^n x^{(i)} = 1\}$ and the l_1 setup is used (see subsection 2.2.4). In this case, subproblems of the form $\min_{x \in Q} \{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_E^2\}$ need $O(n \log(n))$ operations (see section 5.1. in [59]). On the other hand, subproblems of the form $\min_{x \in Q} \{LV(x, z_k) + \alpha_{k+1} \langle \nabla f(x_{k+1}), x - z_k \rangle\}$ are much cheaper since they can be solved in closed-form (see equation 2.14 in subsection 2.2.4).

Furthermore, this modification does not change at all the convergence rate, the Theorem 2.16 remains valid.

Strongly Convex Case

In the smooth strongly convex, the generic denomination fast gradient method denotes any method able to reach the optimal complexity $O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{LR^2}{\epsilon}\right)\right)$ (instead of $O\left(\frac{L}{\mu} \log\left(\frac{LR^2}{\epsilon}\right)\right)$ for the PGM and DGM). Such kind of methods have already been developed (at least with the Euclidean setup) and are typically obtained as a slight modification of existing fast-gradient methods in the non-strongly convex case (see for example [58]).

We develop here a generalization to the strongly convex case of the fast gradient method introduced in [59], that we have studied in the previous subsection in the smooth convex case. As it is the case for all the existing fast gradient

methods for smooth strongly convex problems, we restrict ourselves to the Euclidean setup.

In order to take advantage of the strong convexity, this scheme must be modified:

1. In the choice of the coefficients, that takes into account the existence of a strong convexity parameter $\mu > 0$. The sequence of coefficients must now satisfy the recurrence

$$L + \mu A_k(\mu) = \frac{L\alpha_{k+1}^2}{A_{k+1}(\mu)}, \quad \alpha_0 = 1 \tag{2.26}$$

   where $A_k(\mu) = \sum_{i=0}^k \alpha_i$. (When $\mu = 0$, we have the condition $\alpha_{k+1}^2 = A_{k+1}(0)$, similar to $\alpha_i = \frac{i+1}{2}$.)

2. In the choice of the model which, using the strong convexity of f, becomes

$$\Psi_k(x) = Ld(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2].$$

   (When $\mu = 0$, we retrieve the model (2.25).)

With these modifications, we obtain the following fast gradient method for smooth strongly convex problems:

Algorithm 9 Fast Gradient Method (FGM) for strongly convex problems

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f(x_k), \nabla f(x_k))$.
4: Compute $y_k = T_L(x_k, \nabla f(x_k))$
5: Compute $z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^k \alpha_i [\langle \nabla f(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2]\}$
6: Define $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
7: **end for**

As in the non-strongly convex case, the sequences $\{\Psi_k(x)\}_{k \geq 0}$ and $\{y_k\}_{k \geq 0}$ define a sequence of estimate functions and we obtain the general convergence rate (see Theorems 5.8 and 5.9 in Chapter 5 with $\delta = 0$)

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k(\mu)}.$$

The difference with the non-strongly convex case only comes from the faster growth of the sequence of coefficients $\{\alpha_k\}_{k \geq 0}$. The sequence $\{A_k\}_{k \geq 0}$ defined

by the recurrence (2.26) satisfies (see Theorems 5.10 and 5.14 in Chapter 5 with $\delta = 0$)

$$A_k \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k} \geq \exp\left(\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) \quad \forall k \geq 0.$$

Furthermore, the behavior of the Fast Gradient Method when $\mu > 0$ is never worse than in the case $\mu = 0$. Indeed, we have (see remark 5.11 in Chapter 5)

$$A_k(\mu) \geq A_k(0), \forall k \geq 0$$

and (see remark 5.12 in Chapter 5)

$$A_k(0) \geq \frac{k^2}{4}.$$

We conclude that the fast gradient method designed for smooth strongly convex problems exhibits the following convergence rate:

Theorem 2.17. *Assume that $f \in S^{1,1}_{\mu,L}(Q)$ is endowed with an exact first-order oracle. Then the sequence y_k generated by the Fast Gradient Method satisfies*

$$f(y_k) - f^* \leq Ld(x^*) \min\left(\frac{4}{k^2}, \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right).$$

We obtain, as expected, a method reaching the optimal complexity $O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{LR^2}{\epsilon}\right)\right)$ for smooth strongly convex problems.

2.5 First-order Methods in Nonsmooth Convex Optimization

Let us now have a look at the world of nonsmooth convex optimization. Here also, we are interested in first-order methods i.e. methods developed for solving such kind of problems using a first-order oracle. The main difference with the smooth case is the fact that the oracle cannot provide us anymore a gradient at the current search point x_k since the gradient may not exist.

Hopefully, when the function is convex, we can replace the gradient $\nabla f(x_k)$ by a subgradient of f at x_k, i.e. a vector $g \in E^*$ such that

$$f(y) \geq f(x_k) + \langle g, y - x_k \rangle, \quad \forall y \in E.$$

As we have seen in section 2.1.2, a nonsmooth convex function can have different subgradients at a same point x_k. In the context of nonsmooth convex optimization, a first-order oracle answers at a search point x_k, the value

of f at this point and an arbitrary subgradient $g(x_k)$ of f at this point: $\mathcal{O}(x_k) = (f(x_k), g(x_k))$. A (black-box) first-order method is therefore a method that updates a sequence of search points $\{x_k\}_{k\geq 0}$ using as only information about f, the value of f and a subgradient of f at the different search points.

Before looking at the main family of first-order methods for nonsmooth convex problems, let us look at what we can expect from such kind of methods.

2.5.1 Optimal complexity

Let us start with the weakest level of convexity, assuming that the function is convex but not necessarily strongly convex. We are interested here in lower complexity bound for first-order methods on nonsmooth convex problems.

Convex Case

For a fixed $M < \infty$, we consider $\mathbb{P}_{NC}(M)$, the class of optimization problems of the form $\min_{x\in Q} f(x)$ where Q is any convex set in E, E is any finite-dimensional vector space and f is any function in $NC_M(Q)$ endowed with an exact first-order oracle.

For solving such class of problems, we consider $\mathbb{M}_{NC}(M, R)$ the family of (black-box) first-order methods applicable to any problem in $\mathbb{P}_{NC}(M)$, such that when applied to a problem $\mathcal{P} \in \mathbb{P}_{NC}(M)$, it starts with an initial point x_0 satisfying $dist(x_0, X^*) \leq R$ (where X^* is the optimal solution set of problem $\mathcal{P}$).

The behavior of a black-box first-order method in $\mathbb{M}_{NC}(M, R)$ cannot be arbitrarily good. The best performance that we can expect is given by the following lower-bound on the optimal complexity of first-order methods for nonsmooth convex problems:

Theorem 2.18. *([58, 55]) Using the optimality measure $Opt_f(y) = f(y) - f^*$, the complexity $Compl_{\mathbb{P}_{NC}(M)}(\mathcal{M}, \epsilon)$, of a first-order method $\mathcal{M}$ in $\mathbb{M}_{NC}(M, R)$ on the problem class $\mathbb{P}_{NC}(M)$ cannot be better than*

$$O(1)\frac{M^2R^2}{\epsilon^2}$$

where $O(1)$ denotes an absolute constant factor independent of M, of R and of ϵ.

We conclude that the optimal complexity of the family is such that

$$Compl_{\mathbb{M}_{NC}(M,R),\mathbb{P}_{NC}(M)}(\epsilon) \geq O(1)\frac{M^2R^2}{\epsilon^2}.$$

Furthermore, there exists a family of first-order methods for nonsmooth convex problems (the family of Subgradient/Mirror-descent methods, see [55, 58, 36] and the next subsection) that exhibits such complexity. Therefore, the optimal complexity for first-order methods on nonsmooth convex problems is given by

$$Compl_{\mathbb{M}_{NC}(M,R),\mathbb{P}_{NC}(M)}(\epsilon) = O(1)\frac{M^2R^2}{\epsilon^2}.$$

We see here the big drawback of the lack of smoothness, the best complexity that can be expected is significantly less good than in the smooth case. The difference between $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ and $O\left(\frac{MR}{\epsilon^2}\right)$ is huge even for moderate target accuracy ϵ.

As a consequence of the previous theorem, the best convergence rate that we can expect for a first-order method on nonsmooth convex problems is slow, in $\frac{1}{\sqrt{k}}$ instead of $\frac{1}{k^2}$ in the smooth case:

Theorem 2.19. *[58, 55] The convergence rate $ConvRate_{\mathbb{P}_{NC}(L)}(\mathcal{M},\epsilon)$, of a first-order method $\mathcal{M} \in \mathbb{M}_{NC}(L,R)$, on the problem class $\mathbb{P}_{NC}(L)$, cannot be better than*

$$O(1)\frac{MR}{\sqrt{k}}$$

where $O(1)$ denotes an absolute constant factor independent of M, of R and of k.

Strongly Convex Case

If we assume that the function is strongly convex instead of being only convex, a slightly better complexity can be expected. Let us look at a lower complexity bound for first-order methods on nonsmooth strongly convex problems.

For a fixed $M < \infty$ and a fixed $\mu > 0$, we consider $\mathbb{P}_{NS}(M,\mu)$, the class of optimization problems of the form $\min_{x\in Q} f(x)$ where Q is any convex set in E, E is any finite-dimensional vector space and f is any function in $NS_{\mu,M}(Q)$ endowed with an exact first-order oracle.

Let us define $\mathbb{M}_{NS}(M,\mu,R)$, the family of (black-box) first-order methods applicable to any problem in $\mathbb{P}_{NS}(M,\mu)$ such that when applied to a problem in this class, it uses an initial iterate x_0 satisfying $\|x_0 - x^*\|_E \leq R$.

Exploiting the strong convexity, it is possible to expect a slightly better complexity from a first-order method compared with the nonsmooth, non-strongly convex case :

Theorem 2.20. *([58, 55]) Using the optimality measure $Opt_f(y) = f(y) - f^*$, the complexity $Compl_{\mathbb{P}_{NS}(M,\mu)}(\mathcal{M}, \epsilon)$, of a first-order method $\mathcal{M} \in \mathbb{M}_{NS}(M, \mu, R)$ on the problem class $\mathbb{P}_{NS}(M, \mu)$ cannot be better than*

$$O(1)\frac{M^2}{\mu\epsilon}$$

where $O(1)$ denotes an absolute constant factor independent of M, of μ, of R and of ϵ.

Here also, there exists first-order methods (see for example [36]) reaching this complexity, such that

$$Compl_{\mathbb{M}_{NS}(M,\mu,R),\mathbb{P}_{NS}(M,\mu)}(\epsilon) = O(1)\frac{M^2}{\mu\epsilon}.$$

The strong-convexity allows us to improve the optimal complexity from $O\left(\frac{1}{\epsilon^2}\right)$ to $O\left(\frac{1}{\epsilon}\right)$. However, in comparison to the smooth strongly convex case where a complexity $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$ can be obtained, again the difference is huge.

Using only the optimal complexity bounds, we can already conclude that the lack of smoothness for a function f is clearly a bad news when we want to minimize it using a first-order method. The world of nonsmooth convex optimization is not easy and we cannot expect too much from a first-order method on such kind of problems, at least without additional assumptions on the particular structure of the problem.

2.5.2 Subgradient/ Mirror-descent methods

Let us now look at the most classical family of first-order methods for non-smooth convex problems.

Assume first that we work with an Euclidean setup and that the optimization problem that we want to solve is unconstrained. The principle of the subgradient method is simple, it mimicks the (primal) gradient method, replacing the gradient at the current search point by an arbitrary subgradient at this point

$$x_{k+1} = x_k - \gamma_k g(x_k)$$

where $\{\gamma_k\}_{k\geq 0}$ denotes a sequence of positive stepsizes and $g(x_k) \in \partial f(x_k)$.

This method can be easily generalized to constrained problems. The principle is the same as for the gradient method and we obtain the projected subgradient

method:

$$\begin{aligned} x_{k+1} &= \arg\min_{x\in Q}\{f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\gamma_k}\|x - x_k\|_E^2\} \\ &= \pi_Q\left(x_k - \gamma_k g(x_k)\right). \end{aligned}$$

For a given choice of the stepsizes $\{\gamma_k\}_{k\geq 0}$, the subgradient method becomes

Algorithm 10 Subgradient Method

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $\mathcal{O}(x_k) = (f(x_k), g(x_k))$ where $g(x_k) \in \partial f(x_k)$
4: Compute $x_{k+1} = T_{\frac{1}{\gamma_k}}(x_k, g(x_k)) = \pi_Q\left(x_k - \gamma_k g(x_k)\right)$
5: **end for**

At this point, the subgradient method seems completely similar to the gradient method and we could expect similar behavior. However, due to the nonsmoothness of the objective function, we need to use much smaller stepsizes compared to the smooth case where the coefficients γ_i can be chosen equal to $\frac{1}{L}$. The stepsizes length must be very small from the beginning or must decrease with the iteration. This drawback is a direct consequence of the nonsmoothness of the function, we cannot use too aggressive stepsizes policy in view of the less-predictable behavior of the function.

Theorem 2.21. *[58, 36] Assume that $f \in NC_M(Q)$ is endowed with an exact oracle. Then the sequence $y_k = \arg\min_{x_0,\ldots,x_k} f(x_i)$ or $y_k = \frac{\sum_{i=0}^k \gamma_i x_i}{\sum_{i=0}^k \gamma_i}$, generated using the subgradient method with stepsizes $\{\gamma_i\}_{i\geq 0}$ satisfies*

$$f(y_k) - f^* \leq \frac{\|x_0 - x^*\|_2^2 + M^2\sum_{i=0}^k \gamma_i^2}{2\sum_{i=0}^k \gamma_i} \leq \frac{R^2 + M^2\sum_{i=0}^k \gamma_i^2}{2\sum_{i=0}^k \gamma_i}.$$

We see here that constant stepsizes $\gamma_i = \gamma$ leads to

$$f(y_k) - f^* \leq \frac{R^2}{2(k+1)\gamma} + \frac{M^2\gamma}{2}. \tag{2.27}$$

The first term is exactly the same as what we have obtained for the Primal Gradient Method in the smooth case (where $\gamma = \frac{1}{L}$). The second term comes from the nonsmoothness of the function, and due to this term, the method cannot be convergent when used with constant stepsizes.

If we know a priori the number of iterations that we want to perform, N, we can minimize the upper-bound (2.27) with respect to γ . The optimal constant stepsize is given by $\gamma = \frac{R}{M\sqrt{N+1}}$ for which we have the accuracy guarantee

$$f(y_N) - f^* \leq \frac{MR}{\sqrt{N+1}}.$$

We conclude that the number of iteration N (that we must fix a priori) needed for reaching a target accuracy ϵ is given by

$$N = \frac{M^2R^2}{\epsilon^2} - 1.$$

This complexity has a bad dependence in ϵ but, in view of Theorem 2.18, we cannot expect more for a black-box first-order method on nonsmooth convex problems. The subgradient method is slow but is optimal for a first-order method in $\mathbb{M}_{NC}(M, R)$. This slowness is an intrinsic property of the class $\mathbb{M}_{NC}(M, R)$.

Instead of taking very small stepsizes from the begining $\gamma = \frac{R}{M\sqrt{N+1}}$, we can also use a sequence of decreasing stepsizes which is not based on an a priori knowledge of the number of iterations to perform. Choosing $\gamma_i = \frac{R}{\|g(x_k)\|_E^*\sqrt{i+1}}$, we obtain (see [36])

$$f(y_k) - f^* \leq \frac{MR}{2\sqrt{k+1}}(1 + \ln(k+2)).$$

We obtain in this case a convergent method ($f(y_k) - f^* \to 0$) with, up to a logarithmic factor, optimal convergence rate $O\left(\frac{MR}{\sqrt{k}}\right)$ and therefore optimal complexity $O\left(\frac{M^2R^2}{\epsilon^2}\right)$.

The subgradient method can be easily extended to a non-Euclidean setup. The non-Euclidean version of the subgradient method is known under the name of mirror-descent method (see [36, 3] for example). The principle is completely similar to what we have done for the PGM in the smooth case. The step $x_{k+1} = T_{\frac{1}{\gamma_k}}(x_k, g(x_k))$ is simply replaced by

$$x_{k+1} = \arg\min_{x \in Q}\{f(x_k) + \langle g(x_k), x - x_k\rangle + \frac{1}{\gamma_k}V(x, x_k)\}$$

and the initial point x_0 is chosen as the minimizer of the prox-function d (or the prox-function d is chosen such that x_0 is its minimizer).

The mirror-descent method is therefore given by:

Algorithm 11 Mirror-descent method

1: Compute $x_0 = \arg\min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $\mathcal{O}(x_k) = (f(x_k), g(x_k))$ where $g(x_k) \in \partial f(x_k)$
4: Compute $x_{k+1} = \arg\min_{x \in Q} \{f(x_k) + \langle g(x_k), x - x_k \rangle + \frac{1}{\gamma_k} V(x, x_k)\}$
5: **end for**

and its convergence rate by

Theorem 2.22. *[36] Assume that $f \in NC_M(Q)$ is endowed with an exact oracle. Then the sequence $y_k = \arg\min_{x_0,\ldots,x_k} f(x_i)$ or $y_k = \frac{\sum_{i=0}^{k} \gamma_i x_i}{\sum_{i=0}^{k} \gamma_i}$, generated using the mirror-descent method with stepsizes $\{\gamma_i\}_{i \geq 0}$ satisfies*

$$f(y_k) - f^* \leq \frac{V(x_0, x^*) + \frac{M}{2}^2 \sum_{i=0}^{k} \gamma_i^2}{\sum_{i=0}^{k} \gamma_i}.$$

When used with the Euclidean setup $d(x) = \frac{1}{2} \|x_0 - x\|_2^2$, we retrieve exactly the subgradient method and its convergence rate. All our discussion with respect to the stepsizes choice in the subgradient method remains valid for the mirror-descent method, with R representing now an upper-bound on $\sqrt{2V(x^*, x_0)}$.

Remark 2.29. The Subgradient method/Mirror-descent method can also be adapted to the nonsmooth strongly convex case. This modification, based on multiple restarts of the method with decreasing initial distance to the optimal solution, can reach the optimal complexity $O\left(\frac{M^2}{\mu\epsilon}\right)$ for a first-order method on $\mathbb{P}_{NS}(M, \mu)$. For more details, see section 1.4 in [36].

Remark 2.30. In nonsmooth convex optimization, the easiest scheme (the subgradient method) that we can imagine exhibits already the optimal complexity. However, the practical performance of a first-order method on $\mathcal{P}_{NC}(M)$ can be improved passing to a bundle method (see for example [46, 47, 48]) that uses explicitly both the latest and the previous first-order information, or using a primal-dual scheme (see [63]).

The optimal complexities and corresponding optimal methods for our four classes of convex problems can be summarized as follows:

Class	Optimal Complexity	Optimal Methods
$NC_M(Q)$: f convex Bounded Subgradients	$\Theta\left(\frac{M^2R^2}{\epsilon^2}\right)$	Subgradient/ Mirror Descent Method
$NS_{\mu,M}(Q)$: f Strongly convex Bounded Subgradients	$\Theta\left(\frac{M^2}{\mu\epsilon}\right)$	Subgradient/ Mirror Descent Method
$F_L^{1,1}(Q)$: f convex ∇f Lipschitz-continuous	$\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$	Fast Gradient Method
$S_{\mu,L}^{1,1}(Q)$: f Strong. convex ∇f Lipschitz-continuous	$\Theta\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\epsilon}\right)\right)$	Fast Gradient Method

2.5.3 Using the problem structure: looking inside the black-box

The previous subsections provide us with two bad news: the most natural first-order method for nonsmooth convex problems, the subgradient method, is slow and furthermore, we cannot expect more for a black-box first-order method able to solve any problem in $\mathbb{P}_{NC}(M)$.

The slowness is an intrinsic property of the family of black-box first-order methods $\mathbb{M}_{NC}(M,R)$, it is impossible to obtain a method in this class with a better complexity than $O\left(\frac{M^2R^2}{\epsilon^2}\right)$.

In view of such optimal complexity, it seems therefore hopeless to solve in a reasonable time a nonsmooth convex problem with good target accuracy. This is true if we want to use a first-order method

1. which is 'universal' i.e. applicable to any nonsmooth convex function f with bounded subgradients
2. which is based on a local black-box first-order oracle i.e. a method using only, as information about the particular function f to minimize, the value of f and one of its subgradients at some search points $\{x_k\}_{k\geq 0}$.

If we want to keep such level of generality, there is no hope to do better than the slow subgradient method. However, if we accept

1. to restrict ourselves to a particular subclass of nonsmooth convex problem
2. to use explicitely the structure of the problem (i.e. to look inside the black-box)

it is possible to develop much faster first-order methods.

The use of the problem structure in nonsmooth convex optimization has attracted a lot of interest in the last decade (see for example [59, 60, 61, 4, 62]) and allows us to solve some classes of nonsmooth convex problems efficiently, without being limited by the bad complexity bound proportional to $O\left(\frac{1}{\epsilon^2}\right)$.

We describe here two different situations where, looking inside the black-box, we can use the problem structure to develop first-order methods with better complexity.

Smoothing Technique

Let us consider the following convex optimization problem

$$\min_{x \in Q} f(x) \tag{2.28}$$

where Q is a bounded closed convex set and the function that we want to minimize exhibits the following max-structure

$$f(x) = \max_{u \in U} \{\langle Au, x\rangle - \hat{\phi}(u)\}. \tag{2.29}$$

In this definition, U is a bounded convex set in a finite dimensional space F, $A : F \to E^*$ is a linear operator and $\hat{\phi}(u)$ is a continuous convex function in U.

This function $f : Q \to \mathbb{R}$ is convex but typically nonsmooth. Indeed, using Danskin's Theorem, we have

$$\partial f(x) = \{Au_x^* | \langle Au_x^*, x\rangle - \hat{\phi}(u_x^*) = f(x)\}.$$

As nothing prevents the subproblem defining $f(x)$ at a given x from having multiple optimal solutions, the subdifferential $\partial f(x)$ contains typically more than one element and the function is therefore non-differentiable.

Therefore, if we want to use a black-box first-order method for nonsmooth convex problems, like the subgradient method, we cannot expect to solve the

optimization problem (2.28) with a better complexity than $O\left(\frac{1}{\epsilon^2}\right)$.

However, using the particular max-structure of f, it is possible to develop a first-order method with complexity $O\left(\frac{1}{\epsilon}\right)$.

First, let us construct a smooth approximation of f using a prox-function $d_U(.)$ on U and a strictly positive parameter μ:

$$f_\mu(x) = \max_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u) - \mu d_U(u)\}.$$

With this modification, the subproblem defining $f_\mu(x)$ at a given x (i.e. $\max_{u\in U}\{\langle Au, x\rangle - \hat{\phi}(u) - \mu d_U(u)\}$), becomes a strongly concave problem in u. This subproblem has therefore only one optimal solution u_x^*. We conclude that $\partial f_\mu(x)$ contains one and only one element for all $x \in E$ and the function $f_\mu(.)$ is therefore differentiable.

More precisely, we have that the function f_μ has a Lipschtz-continuous gradient (see [59])

$$\nabla f_\mu(x) = Au_x^*, \quad \text{where } u_x^* = \arg\max_{u\in U}\{\langle Au, x\rangle - \hat{\phi}(u) - \mu d_U(u)\}$$

with constant

$$L_\mu = \frac{1}{\mu}\|A\|_{F,E^*}^2$$

where $\|A\|_{F,E^*} = \max_{x,u}\{\langle Au, x\rangle, | \|x\|_E = 1, \|u\|_F = 1\}$.

Furthermore, the smooth function f_μ is a good approximation of the original nonsmooth function f (see [59]) in the following sense

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D_U \tag{2.30}$$

where $D_U = \max\{d_U(u) : u \in U\}$.

Therefore, in order to find an ϵ-solution for the original problem (2.28), we can apply the fast-gradient method to a smooth approximation $f_\mu(x)$. It remains to choose the value of the parameter μ and the number of iterations of the FGM that we have to perform.

In view of Theorem 2.16, if we apply N iterations of the fast gradient method to the function $f_\mu(x)$, we obtain an approximate solution y_N such that

$$f_\mu(y_N) - f_\mu^* \leq \frac{4L_\mu D_Q}{(N+1)^2}$$

where $f_\mu^* = \min_{x\in Q} f_\mu(x)$, $D_Q = \max_{x\in Q} d_Q(x)$ and $d_Q(.)$ is a prox-function on Q. Therefore in view of inequality (2.30), we have

$$f(x) - f^* \leq \frac{4L_\mu D_Q}{(N+1)^2} + D_U\mu.$$

As we are looking for an ϵ-solution for the optimization problem (2.28), we can

1. Choose μ such that $D_U\mu = \frac{\epsilon}{2}$ i.e. $\mu = \frac{\epsilon}{2D_U}$
2. Choose N such that $\frac{4L_\mu D_Q}{(N+1)^2} = \frac{\epsilon}{2}$. As $L_\mu = \frac{1}{\mu}\|A\|^2_{F,E^*} = \frac{2D_U\|A\|^2_{F,E^*}}{\epsilon}$, this conditions gives us

$$N = \frac{4(D_Q D_U)^{1/2}\|A\|_{F,E^*}}{\epsilon} - 1.$$

It means that using the smoothing technique, we are able to solve, up to target accuracy $\epsilon > 0$, the nonsmooth convex problem (2.28) with max-structure (2.29) in $O\left(\frac{1}{\epsilon}\right)$ iterations (instead of $O\left(\frac{1}{\epsilon^2}\right)$ with a black-box subgradient method). We see here clearly the interest to use explicitly the problem structure in the scheme.

Remark 2.31. Seeing the problem (2.28) with max-structure (2.29) as a convex-concave saddle point problem $\min_{x\in Q}\max_{u\in U}\Phi(x,u)$ where $\Phi(.,.)$ has a Lipschitz-continuous gradient, it is also possible to obtain the complexity $O\left(\frac{1}{\epsilon}\right)$ using the mirror-prox method (see [51, 37]).

Composite Objective function

Let us now consider another case of nonsmooth convex problems where the problem structure can be used to obtain an efficient first-order method. Assume that the objective function that we want to minimize on a closed convex set Q has the following composite structure

$$f(x) = f_1(x) + f_2(x)$$

where

- $\diamond$ f_1 is a smooth convex function belonging to $F_L^{1,1}(Q)$ endowed with a black-box first-order oracle
- $\diamond$ f_2 is a closed convex function (typically nonsmooth) assumed to be easy.

For a particular setup choice, the fact that f_2 is easy means that for all $g \in E^*$, subproblems of the form

$$\min_{x\in Q}\{\langle g, x\rangle + Cd(x) + f_2(x)\},$$

with $C > 0$, are easy to solve.

Remark 2.32. In this case, as $V(x,z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, for all $z \in E$ and $g \in E^*$ and $C > 0$, subproblems of the form

$$\min_{x \in Q} \{\langle g, x \rangle + CV(x,z) + f_2(x)\},$$

are also easy to solve.

When f_2 is nonsmooth, the whole function f can be seen as a nonsmooth convex function endowed with a first-order oracle. We could therefore minimize it using the subgradient/mirror-descent method. However such a black-box approach suffers from a bad complexity proportional to $O\left(\frac{1}{\epsilon^2}\right)$ and does not use the particular structure of f, in particular the fact that f_2 is assumed to be easy.

Another approach can be used. Developed in [62, 4], it consists of applying the methods of smooth convex optimization like the PGM, the DGM or the FGM to the composite function $f(x) = f_1(x) + f_2(x)$ but with f_2 kept intact in the subproblems.

More precisely, in these three methods, subproblems of the form

$$\arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle]\},$$

$$\arg\min_{x \in Q} \{LV(x, x_k) + \langle \nabla f(x_k), x - x_k \rangle\}$$

becomes respectively

$$\arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f_1(x_i) + \langle \nabla f_1(x_i), x - x_i \rangle] + A_k f_2(x)\} \tag{2.31}$$

$$\arg\min_{x \in Q} \{LV(x, x_k) + \langle \nabla f_1(x_k), x - x_k \rangle + f_2(x)\} \tag{2.32}$$

Furthermore, the presence of f_2 does not affect the complexity of these methods i.e. $O\left(\frac{LR^2}{\epsilon}\right)$ for the Primal and Dual Gradient Methods and $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ for the Fast Gradient Method.

It means that, even if the whole function f is nonsmooth (since f_2 is nonsmooth), we can solve the problem as efficiently as a smooth problem. However, such approach is possible only when the nonsmooth part f_2 is sufficiently easy,

otherwise the subproblems (2.31), (2.32) become intractable and even one iteration cannot be performed.

An important example is the l_1-regularized problem

$$\min_{x\in\mathbb{R}^n}\{f_1(x)+\lambda\|x\|_1\} \tag{2.33}$$

where $f_1 \in F_L^{1,1}(\mathbb{R}^n)$ and $\lambda > 0$ is a regularization parameter. Using the Euclidean setup, the nonsmooth convex function $f_2(x) = \lambda\|x\|_1$ can be seen as an easy function. Indeed, we have

$$\begin{aligned} w_k &= \arg\min_{x\in Q}\{\langle\nabla f_1(x_k), x-x_k\rangle + \frac{L}{2}\|x-x_k\|_2^2 + \lambda\|x\|_1\} \\ &= \arg\min_{x\in Q}\{\frac{L}{2}\left\|x-(x_k-\frac{1}{L}\nabla f_1(x_k))\right\|_2^2 + \lambda\|x\|_1\} \\ &= \mathcal{T}_{\frac{\lambda}{L}}(x_k-\frac{1}{L}\nabla f_1(x_k)) \end{aligned}$$

where $\mathcal{T}_\alpha(x)_i = (|x_i|-\alpha)_+\text{sgn}(x_i)$ is the shrinkage operator (see [4]). Similarly, we have

$$\begin{aligned} z_k &= \arg\min_{x\in Q}\{\langle\sum_{i=0}^k \alpha_i\nabla f_1(x_i), x\rangle + \frac{L}{2}\|x-x_0\|^2 + A_k\lambda\|x\|_1\} \\ &= \mathcal{T}_{\frac{A_k\lambda}{L}}(x_0 - \frac{\sum_{i=0}^k \alpha_i\nabla f_1(x_i)}{L}). \end{aligned}$$

The subproblems of the PGM, DGM and FGM can be solved in closed-form, the presence of the l_1 term does not affect the (analytical) complexity of these methods. Even if the objective function $f(x) = f_1(x) + \lambda\|x\|_1$ is nonsmooth, an ϵ-solution for the problem 2.33 can be obtained after only $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations using the Fast Gradient Method.

Remark 2.33. In [62, 4], this efficient approach for composite convex problems has been developed using an Euclidean setup. These results can be easily extended to a non-Euclidean setup as proved in Chapter 7 of this thesis.

2.6 First-order Methods beyond their comfort zone

In this thesis, we want to extend the scope of the first-order methods beyond the comfort zone described in this chapter. The three challenging difficulties

we consider in this thesis are inexactness in the first-order information, lack of smoothness for the objective function and presence of linear constraints (making computationally difficult the projections on the feasible set).

The main contributions of this thesis in these three directions can be summarized as the answers (that will be given chapter after chapter) to the following questions:

1. **Inexactness in first-order information:**

   - ◇ What is the effect of (deterministic or stochastic) inexact first-order information on the existing first-order methods of smooth convex optimization, namely the Primal Gradient Method (PGM), the Dual Gradient Method (DGM) and the Fast Gradient Method (FGM) ? Answers in Chapters 4, 5 and 7.
   - ◇ Is it possible to develop a first-order method which is a the same time fast and robust with respect to oracle errors ? Or is there an intrinsic link between the speed of convergence of a method and its sensitivity with respect to errors ? Answer in Chapters 4 and 5.
   - ◇ Can we exploit the strong convexity of an objective function in order to reduce the sensitivity of first-order methods with respect to oracle errors ? Answer in Chapter 5.
   - ◇ Is it possible to develop new methods with optimized behavior in view of the level of errors in first-order information ? Answer in Chapter 6.
   - ◇ Is the case of a stochastic inexact oracle in some sense more favorable compared to the deterministic one ? Answer in Chapter 7.
   - ◇ Is it possible to modify existing first-order methods of smooth convex optimization in order to reduce the effect of a stochastic noise to zero ? At which rate ? Answer in Chapter 7.

2. **Lack of smoothness:**

   - ◇ Is it possible to apply first-order methods, initially designed for smooth convex problems, to objective functions with a weaker level of smoothness ? Answer in Chapter 4.
   - ◇ What is the behavior of the PGM, DGM and FGM when applied to a nonsmooth or weakly smooth function ? How to choose the stepsizes, taking into account the lack of smoothness, in a optimal way ? Answer in Chapter 4.

- ◇ Is it possible to obtain a universal optimal first-order method, exhibiting an optimal complexity on smooth, weakly smooth and non-smooth convex problems ? Answer in Chapter 4.
- ◇ Is it possible to apply first-order methods, initially designed for smooth strongly convex problems, to strongly convex functions with weaker level of smoothness or with different level of convexity? What is the complexity of the PGM, DGM and FGM on these different classes of problems? Answer in Chapter 5.

3. **Presence of linear constraints:**

- ◇ How to apply a first-order method to a linearly constrained problem for which projections on the feasible set are intractable ? Answer in Chapter 3.
- ◇ Why isn't a simple dualization of the linear constraints sufficient to obtain an efficient approach ? Answer in Chapter 3.
- ◇ Is the smoothing technique useful in order to improve the dual convergence rate ? Answer in Chapter 3.
- ◇ How to reconstruct a nearly optimal and nearly feasible primal solution from a nearly optimal dual solution ? Answer in Chapter 3.
- ◇ Why is a double smoothing of the dual function needed to have the same fast convergence for the dual and the primal processes ? Answer in Chapter 3.

Chapter 3

Double Smoothing Technique for Linearly Constrained Convex Optimization Problems

This chapter corresponds to the paper [23]:
O. Devolder, F. Glineur and Y. Nesterov. **Double Smoothing Technique for Large-Scale Linearly Constrained Convex Optimization**. *SIAM Journal of Optimization, Volume 22, Issue 2*, (2012)

Chapter 3 in four Questions/Answers.

- ◇ *How to apply a first-order method to a linearly constrained problem for which projections on the feasible set are intractable ?*

  A classical approach consists of dualizing the linear constraints, obtaining an uconstrained dual problem. This dual approach is particularly interesting in the situation where the linear constraints are the only coupling constraints between the variables. In this case, after dualization of the linear constraints, the dual objective function typically becomes separable and therefore easy to compute.

- ◇ *Why is a simple dualization of the linear constraints not sufficient in order to obtain an efficient approach ?*

  After dualization of the linear constraints, we obtain an unconstrained dual problem that is typically non differentiable. We can therefore only apply to this dual function a first-order method of nonsmooth convex optimization, such as the subgradient method, and obtain an unattractive complexity of order $O\left(\frac{1}{\epsilon^2}\right)$ both for the primal and dual convergence.

◇ *Is the smoothing technique useful in order to improve the dual convergence rate ?*

Using the structure of the dual nonsmooth function, it is possible to apply the smoothing technique developed by Nesterov in [59], that transforms the nonsmooth convex dual function into a smooth convex approximation and applies to this function a fast gradient method. This approach allows us to solve the dual problem with a significantly better complexity proportional to $O(\frac{1}{\epsilon})$.

◇ *Why is a double smoothing of the dual function needed ?*

The complexity improvement obtained with the classical smoothing technique is lost during the process of reconstructing a primal solution from the dual one. Because we cannot guarantee a fast decrease of the norm of the gradient, primal convergence (the one in which we are really interested) reduces to a complexity not better than using the basic subgradient approach.

The double smoothing technique consists in going one step further in the smoothing process, making the dual objective not only smooth with a Lipschitz-continuous gradient but also strongly convex. Applying a fast gradient method for smooth strongly convex functions to this doubly smoothed dual objective function, we can solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations and from this nearly optimal dual solution, reconstruct a nearly optimal primal solution with the same level of accuracy.

Contents

3.1 Subgradient method vs Smoothing techniques **69**
3.2 Problem formulation and dual approach **71**
3.3 Examples of problems with separable structure **73**
3.3.1 A finite-dimensional example 73
3.3.2 An infinite-dimensional example 74
3.4 Double Smoothing Technique **74**
3.4.1 First Smoothing 75
3.4.2 Second Smoothing 78
3.5 Strong duality and norm of dual optimal solutions . . . **79**
3.6 Solving the primal-dual problem **82**
3.6.1 Convergence of $\theta(z_k)$ to θ^* 82
3.6.2 Convergence of $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}$ 84
3.6.3 Constructing an approximate primal solution . . 85

3.7 Practical implementation **86**
3.8 Application in Optimal Control **88**
3.8.1 Class of optimal control problems and reformulation 90
3.8.2 Evaluation of $\|\mathcal{A}_i\|_2$ 92
3.8.3 Bounding the growth of norms $\|\mathcal{A}_i\|_2$ with time 94
3.9 Comparison with the literature and conclusion **96**

In large-scale convex optimization, first-order methods are often the methods of choice due to their cheap iteration cost. In particular, constrained problems can be solved provided that performing projection on their feasible set is computationally easy. In this chapter, we assume that both the convex objective function J, defined on the Hilbert space U, and the convex feasible region $S \subset U$ are sufficiently simple so that the problem $\min_{u \in S} J(u)$ can be solved efficiently, or even in closed-form. However, the situation becomes completely different when adding the constraint $\mathcal{A}u \in T$, based on a linear operator $\mathcal{A} : U \to V^*$, where V is a finite-dimensional Euclidean space and T a bounded closed convex set in V^* (the dual space of V). Indeed, the problem may become difficult because projection onto the new feasible set $\{u \in S : \mathcal{A}u \in T\}$ may be computationally very expensive, or even intractable.

A natural approach is therefore to dualize this difficult linear constraint, obtaining the primal-dual pair of problems

$$P^* = \min_{u \in S} \{J(u) + \max_{z \in V} [\langle \mathcal{A}u, z \rangle - \sigma_T(z)]\}$$

$$D^* = \max_{z \in V} \{-\sigma_T(z) + \min_{u \in S} [J(u) + \langle \mathcal{A}u, z \rangle]\}$$

where $\sigma_T(z) = \sup_{x \in T} \langle x, z \rangle$ denotes the support function of set T, defined on V.

In this chapter, we assume that the dimension of set V (i.e. the size of the linear constraints) is small compared to the dimension of set U, the latter being allowed to be infinite. Due to this asymmetry, we are led to consider a purely dual algorithmic scheme, generating its iterates only in the low-dimensional space V. The only operation we need to be able to perform in the infinite-dimensional or high-dimensional space U is the computation of the value of dual objective function at a given point $z \in V$, which requires solving the optimization subproblem $\min_{u \in S} J(u) + \langle \mathcal{A}u, z \rangle$ over the simple set S.

Example 3.1. As a first motivation, consider the purely linear infinite-dimensional problem:

$$P^* = \inf_{u \in L^2([\alpha,\beta]) : M_1 \leq u(t) \leq M_2} \int_\alpha^\beta f(t)u(t)dt : \int_\alpha^\beta a_i(t)u(t)dt = b_i \forall i = 1, \ldots, m$$

where we seek a function u in Hilbert space $L^2([\alpha,\beta])$ and data consists of functions a_i $(1 \leq i \leq m)$ and f in $L^2([\alpha,\beta])$. Since we are working in L^2, inequalities on $u(t)$ are to be understood as to hold almost everywhere. It is straightforward to define

$$U = L^2([\alpha,\beta]), \quad S = \{u \in U : M_1 \leq u(t) \leq M_2, t \in [\alpha,\beta]\},$$
$$J(u) = \int_\alpha^\beta f(t)u(t)dt, \quad V = R^m, \quad T = \{b\} \subset V,$$
$$\mathcal{A} : U \to V^* : u \to \left(\int_\alpha^\beta a_1(t)u(t)dt, \ldots, \int_\alpha^\beta a_m(t)u(t)dt \right)^T$$

so that the problem fits our formulation $\min_{u\in S} J(u) : \mathcal{A}u \in T$. Dualizing the linear equality constraints, we obtain the dual function:

$$\Theta(z) = -\sum_{i=1}^m z_i b_i + \inf_{u\in S} \left[\int_\alpha^\beta f(t)u(t)dt + \sum_{i=1}^m \left(\int_\alpha^\beta a_i(t)u(t)dt \right) z_i \right] \quad (3.1)$$
$$= -\sum_{i=1}^m z_i b_i + \inf_{u\in S} \int_\alpha^\beta \Big(f(t) + \sum_{i=1}^m a_i(t)z_i \Big) u(t)dt. \quad (3.2)$$

Due to the fact that only pointwise constraints $M_1 \leq u(t) \leq M_2$ are still present in this problem, we can solve it in a pointwise way, minimizing u for each value of t separately. Indeed, any solution $u_z \in S$ satisfying $u_z(t) = M_1$ when $f(t) + \sum_{i=1}^m a_i(t)z_i > 0$ and $u_z(t) = M_2$ when $f(t) + \sum_{i=1}^m a_i(t)z_i < 0$ is optimal for problem (3.1).
Since we are able to compute the value of $\Theta(z)$ in closed form for any value of z, we can apply a first-order method to the finite-dimensional problem $\max_{z\in V} \Theta(z)$.

Our goal in this work is to show that it is possible to solve the dual problem efficiently and reconstruct from this process a nearly optimal and feasible primal solution. We develop to that effect a new double smoothing approach, which is a variant of the smoothing techniques described in subsection 2.5.3 and in [59, 60, 61]. This technique uses the problem structure to regularize the dual objective function into a smooth strongly convex function with Lipschitz continuous gradient. These modifications allow us to minimize the dual function with an optimal gradient scheme in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$iterations, where ϵ is the desired accuracy. From the dual minimization sequence, we reconstruct a nearly feasible and optimal primal solution, whose accuracy can be controlled by parameters of our algorithm.

The structure of this chapter is as follows. In the Section 3.1, we recall briefly two standard approaches for solving a nonsmooth convex optimization problem

with a first-order method: subgradient-type schemes and smoothing techniques. We also present the first-order methods that can be used to solve efficiently the smoothed problem obtained by the smoothing technique. In particular, we recall the optimal method [58] for smooth and strongly convex functions and describe its rate of convergence. In Section 3.2, we present in a more general form our problem class and derive the corresponding dual problem. Using Danskin's Theorem, we show that the dual objective function is in general nonsmooth. Section 3.3 presents two simple examples of problems (one finite-dimensional, the other infinite-dimensional) with separable structure that fit our problem class. The double smoothing is described in Section 3.4, where we apply two regularizations to the dual objective function in order to make it smooth and strongly convex. We also explain the necessity of requiring both properties. In Section 3.5, we study under which regularity conditions strong duality holds and how it is possible to bound the size of the dual optimal set. This bound will be useful in the convergence analysis of our scheme. In Section 3.6, an optimal first-order method is applied to the modified dual objective function and a nearly feasible and optimal primal solution is reconstructed from the dual minimization sequence. Accuracy of the primal and dual solutions can be adjusted by parameters of our algorithm. In Section 3.7, we discuss the practical implementation of our method, in particular when the size of the dual optimal set is unknown. In Section 3.8, we consider applications of our double smoothing technique to optimal control problems. We conclude this chapter with a comparison between our results and the existing literature.

3.1 Subgradient method vs Smoothing techniques

Consider the convex optimization problem $\min_{y \in V} f(y)$ where $f : V \to R$ is a convex function defined on the finite-dimensional space V. If f is non-differentiable, we know that the complexity of a black-box first-order method that does not use the problem structure cannot be better that $O\left(\frac{1}{\epsilon^2}\right)$ iterations, where ϵ is the desired accuracy for the objective function (see [55, 58]). This lower bound is achievable by various first-order methods for nonsmooth convex problems, such as subgradient methods (see e.g. [58, 63]). These schemes can therefore be applied directly to a nonsmooth convex function, albeit with a relatively slow convergence rate.

When the nonsmooth function has a particular saddle-point structure:

$$f(y) = \max_{u \in S} \{g(u) + \langle Au, y \rangle\} \tag{3.3}$$

where $g : U \to R$ is concave on the finite-dimensional space U and $S \subset U$ is closed and convex, another approach can be used. In the smoothing technique developed in [59, 60, 61], this nonsmooth function is approximated by a

smooth one and an optimal first-order method of smooth convex optimization is applied to the smooth approximation. With this approach, we can solve the original nonsmooth problem up to accuracy ϵ in only $O\left(\frac{1}{\epsilon}\right)$ iterations (instead of $O\left(\frac{1}{\epsilon^2}\right)$ with a subgradient scheme).

In the smoothing approach, we want to minimize the original nonsmooth function with an accuracy ϵ. We construct a smooth approximation of f belonging to $F_{L(\epsilon)}^{1,1}$ with constant $L(\epsilon) = \Theta(\frac{1}{\epsilon})$. Applying a fast gradient method for $F_{L(\epsilon)}^{1,1}$ to this smoothed function, we can solve the original problem with the desired accuracy in $O(\frac{1}{\epsilon})$ iterations. In our work, the function that we optimize using a first-order method, with saddle-point structure (3.3), is the dual objective function. However, our goal is to solve efficiently the primal problem, not the dual one.

In order to reconstruct a good primal solution from the iterates of the numerical scheme applied to the dual problem, we will need to apply a second smoothing to the dual function before we apply the fast gradient method. Its purpose will be to ensure strong convexity of the resulting dual objective function. Fast gradient methods that are optimal for smooth strongly convex objective functions are also known (see for example subsection 2.4.5 and [58]), and we will use such a method to minimize the doubly smoothed dual objective function.

For the reader's convenience, we conclude this section with a presentation of the simplest optimal method for minimizing smooth strongly convex functions on the whole space (see subsection 2.2.1 in [58]). Let function $f \in S_{\kappa,L}^{1,1}(V)$ and consider the problem: $\min_{y \in V} f(y)$. We assume that this problem is solvable. Denote by f^* its optimal value and by y^* the optimal solution.

Algorithm 12 Simple Fast Gradient Method (FGM) for $S_{\kappa,L}^{1,1}(V)$

1: Choose $w_0 = y_0 \in V$
2: **for** k = 0 : ... **do**
3: Compute $y_{k+1} = w_k - \frac{1}{L}\nabla f(w_k)$
4: Compute $w_{k+1} = y_{k+1} + \frac{\sqrt{L}-\sqrt{\kappa}}{\sqrt{L}+\sqrt{\kappa}}(y_{k+1} - y_k)$.
5: **end for**

By Theorem 2.2.3 in [58] we have

$$f(y_k) - f^* \leq \left(f(y_0) - f^* + \frac{\kappa}{2}\|y_0 - y^*\|_V^2\right) e^{-k\sqrt{\frac{\kappa}{L}}} \leq 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}. \tag{3.4}$$

Since ∇f is Lipschitz-continuous, in view of Theorem 2.1.5 in [58] we have

$$\tfrac{1}{2L}\|\nabla f(y_k)\|_{V^*}^2 \quad \leq \quad f(y_k) - f^* \quad \overset{(3.4)}{\leq} \quad 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}.$$

Therefore,

$$\|\nabla f(y_k)\|_{V^*}^2 \quad \leq \quad 4L(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}. \tag{3.5}$$

Finally, since f is strongly convex, by Theorem 2.1.8 in [58] we have

$$\tfrac{\kappa}{2}\|y_k - y^*\|_V^2 \quad \leq \quad f(y_k) - f^* \quad \overset{(3.4)}{\leq} \quad 2(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}.$$

Using this inequality and additional arguments, we conclude that

$$\|y_k - y^*\|_V^2 \quad \leq \quad \min\left\{\|y_0 - y^*\|_V^2, \tfrac{4}{\kappa}(f(y_0) - f^*)e^{-k\sqrt{\frac{\kappa}{L}}}\right\}. \tag{3.6}$$

3.2 Problem formulation and dual approach

As described in the introduction, we consider in this work optimization problems of the form:

$$P^* = \inf_{u \in S} J(u) : \mathcal{A}u \in T. \tag{3.7}$$

where U is a Hilbert space endowed with the Euclidean norm $\|\cdot\|_U = \sqrt{(\cdot|\cdot)_U}$, S is a bounded, closed, convex set in U, V is a finite-dimensional Hilbert space endowed with the Euclidean norm $\|\cdot\|_V = \sqrt{(\cdot|\cdot)_V}$, T is a bounded, closed, convex set in V^*, the dual space of V, $J : U \to R$ is a closed and convex function defined on S and $\mathcal{A} : U \to V^*$ is a bounded linear operator. Space U is allowed to be infinite-dimensional, but the approach used in this chapter is also efficient for large-scale finite-dimensional problems, i.e. when $\dim U \gg \dim V$.

Remark 3.1. Note that problems with multiple linear constraints also belong to problem class (3.7):

$$P^* = \inf_{u \in S} J(u) : \mathcal{A}_i u \in T_i \quad \forall i = 1, \ldots, m. \tag{3.8}$$

Indeed, assume that $V = V_1 \times V_2 \times \cdots \times V_m$ where V_i is a finite-dimensional Hilbert space for $i = 1, \ldots, m$. For each i, $\mathcal{A}_i : U \to V_i^*$ is a bounded linear operator. Let $T = T_1 \times T_2 \times \cdots \times T_m$ where T_i is a bounded closed convex set in V_i^*. Defining the linear operator $\mathcal{A} : U \to V^*$ such that $\langle \mathcal{A}u, z\rangle = \sum_{i=1}^m \langle \mathcal{A}_i u, z_i\rangle$ for every $u \in U$ and every $z = (z_1, ..., z_m) \in V$, the constraint $\mathcal{A}u \in T$ is clearly equivalent to $\mathcal{A}_i u \in T_i \quad \forall i = 1, \ldots, m$. Finally, we have:

$$\begin{aligned}\|\mathcal{A}\| \quad &= \quad \max_{\|u\|_U = 1, \|z\|_V = 1} \langle \mathcal{A}u, z\rangle = \max_{\|u\|_U = 1, \sum_{i=1}^m \|z_i\|_{V_i}^2 = 1} \sum_{i=1}^m \langle \mathcal{A}_i u, z_i\rangle \\ &= \quad \max_{\|u\|_U = 1} \Big[\sum_{i=1}^m \|\mathcal{A}_i u\|_{V_i^*}^2\Big]^{1/2} \leq \Big[\sum_{i=1}^m \|\mathcal{A}_i\|^2\Big]^{1/2}.\end{aligned}$$

Our assumptions on J, S and T are motivated by the following classical result (see for example [78]):

Theorem 3.1. *If X is a reflexive Banach space, $M \subset X$ is a bounded, closed, convex set and $F : X \to R$ is a closed, convex function, then optimization problem $\min_{x \in M} F(x)$ is solvable. Furthermore, if in addition X is an Hilbert space and F is strictly convex, then the optimal solution of this problem is unique.*

We conclude that subproblems of the forms

$$\inf_{u \in S} \{J(u) + \langle \mathcal{A}u, z \rangle\} \quad \text{and} \quad \inf_{x \in T} \langle x, z \rangle$$

are solvable for each $z \in V$ and that subproblems of the form:

$$\inf_{u \in S} \{J(u) + \langle \mathcal{A}u, z \rangle + \frac{\mu}{2} \|u\|_U^2\} \quad \text{and} \quad \inf_{x \in T} \{\langle x, z \rangle + \frac{\rho}{2} \|x\|_{V^*}^2\}$$

have a unique optimal solution for every $z \in V$, $\mu > 0$ and $\rho > 0$.

In our setting, it is natural to dualize the linear constraint $\mathcal{A}u \in T$ and to consider a dual method, working only in the small-dimensional space V. Since T is a closed convex set, inclusion $\mathcal{A}u \in T$ is equivalent to $\langle \mathcal{A}u, z \rangle \leq \sigma_T(z)\ \forall z \in V$, where $\sigma_T(z) = \sup_{x \in T} \langle x, z \rangle$ denotes the support function of T, which allows us to dualize the linear constraints. We obtain the primal-dual pair of problems:

$$P^* = \inf_{u \in S} [J(u) + \sup_{z \in V} (\langle \mathcal{A}u, z \rangle - \sigma_T(z))], \; D^* = \sup_{z \in V} [-\sigma_T(z) + \inf_{u \in S} (J(u) + \langle \mathcal{A}u, z \rangle)].$$

Thus, the Lagrangian dual problem (in minimization form) is given by

$$-D^* = \theta^* = \inf_{z \in V} [\sigma_T(z) + \phi(z)] := \inf_{z \in V} \theta(z) \geq -P^* \tag{3.9}$$

where we define $\phi(z) = \sup_{u \in S} [-J(u) - \langle \mathcal{A}u, z \rangle]$, we recall $\sigma_T(z) = \sup_{x \in T} \langle x, z \rangle$ and we let $\theta(z) = \sigma_T(z) + \phi(z)$.

We made the initial assumption that it is easy to optimize a function over simple set S, so that it is easy to compute the value of $\theta(z)$ for every $z \in V$; however this function is typically non-differentiable. Indeed, using Danskin's Theorem ([22, 8]), the subdifferentials of σ_T and ϕ are given by

$$\begin{aligned} \partial \sigma_T(z) &= \{x \in T : \langle x, z \rangle = \sigma_T(z)\} \\ \partial \phi(z) &= \{-\mathcal{A}u : -J(u) - \langle \mathcal{A}u, z \rangle = \phi(z), u \in T\}. \end{aligned}$$

As the optimization problems defining $\sigma_T(z)$ and $\phi(z)$ can have multiple optimal solution, the above subdifferentials may contain several elements and

function $\theta(z)$ can be nonsmooth. Dualization of problem (3.7) therefore results in a nonsmooth convex problem.

As explained in the previous section, instead of relying on subgradient-type schemes with relatively slow convergence, we will solve the dual problem with a smoothing technique [59, 60, 61]. In the smoothing approach, using the specific structure of the problem, we apply some regularization to the objective function and obtain much faster methods (which are not pure black-box schemes anymore). We develop in this chapter an algorithm able to solve the dual problem with accuracy ϵ and to reconstruct, from a nearly optimal dual solution, a nearly optimal and feasible primal solution in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ iterations.

3.3 Examples of problems with separable structure

Before we go into the details of the double smoothing, we provide two examples of problems with separable structure (one finite-dimensional, the other infinite-dimensional) that belong to our problems class.

3.3.1 A finite-dimensional example

Consider the case where

- $U = R^N = R^{N_1} \times R^{N_2} \times ... \times R^{N_m}$ and $S = S_1 \times S_2 \times ... \times S_m$ where $S_i \subset R^{N_i}$,
- $V = R^n$ whith $n \ll N$ and T is a bounded closed convex set in R^n,
- $J(u) = \sum_{i=1}^m J_i(u_i)$ where $u_i \in R^{N_i}$ and $J_i : R^{N_i} \to R$ is a closed and convex function,
- $\mathcal{A} = (A_1\ A_2\ \cdots\ A_m) \in R^{n\times N}$ where $A_i \in R^{n\times N_i}$.

Our problem becomes $\min \sum_{i=1}^m J_i(u_i) : \sum_{i=1}^m A_i u_i \in T, u_i \in S_i\ \forall i = 1, ..., m$. This problem has a specific structure that we want to exploit. When the coupling constraint $\sum_{i=1}^m A_i u_i \in T$ is dropped, we obtain a separable problem that we can solve separately for each u_i. With this property, it seems natural to dualize the coupling constraint and to consider the dual problem: $\min_{z\in R^n} \theta(z) = \min_{z\in R^n} \sigma_T(z) + \phi(z)$ in the small-dimensional space R^n. For each $z \in R^n$, the dual objective function can be computed in a pointwise way: solving the subproblem $\max_{u\in S}\{-J(u) - \langle Au, z\rangle\}$ is equivalent to solving separately for each i the subproblems $\max_{u_i\in S_i}\{-J_i(u_i) - \langle A_i u_i, z\rangle\}$, which

we assume to be computationally easy. The modified dual objective function obtained after smoothing $\max_{u\in S}\{-J(u) - \langle \mathcal{A}u, z\rangle - \frac{\mu}{2}\|u\|^2_{R^N}\}$ is similarly easy to compute with independent subproblems, since squared Euclidean norm $\|u\|_{R^N}$ is also separable: $\|u\|^2_{R^N} = \sum_{i=1}^m \|u_i\|^2_{R^{N_i}}$.

3.3.2 An infinite-dimensional example

Consider the case where:

- $\diamond$ $U = L^2([0,\mathcal{T}], R^m)$ and $S = \{u \in U : u(t) \in S(t)\ \forall t\}$, where $S(t)$ is a closed, convex set in R^m for each $t \in [0,\mathcal{T}]$ and $\cup_{t\in[0,\mathcal{T}]} S(t)$ is bounded,
- $\diamond$ $V = R^n$ and T is a bounded, closed, convex set in R^n,
- $\diamond$ $J(u) = \int_0^{\mathcal{T}} F(t, u(t))dt$ where function $F : [0,\mathcal{T}] \times R^m \to R$ is convex and continuous,
- $\diamond$ $\mathcal{A} : U \to R^n$ is defined by $\mathcal{A}u = \int_0^{\mathcal{T}} A(t)u(t)dt$ where $\int_0^{\mathcal{T}} \|A(t)\|_2^2\, dt < +\infty$ and $A(t) \in R^{n\times m}\ \forall t \in [0,\mathcal{T}]$.

Our problem becomes:

$$\inf \int_0^{\mathcal{T}} F(t, u(t))dt : \int_0^{\mathcal{T}} A(t)u(t)dt \in T, u(t) \in S(t) \quad \forall t \in [0,\mathcal{T}]. \quad (3.10)$$

When the linear coupling constraint is dropped, we obtain a separable problem that we can solve independently for each $t \in [0,\mathcal{T}]$: $\min_{u(t)\in S(t)} F(t, u(t))$. Hence, dualization of the linear coupling constraint is here also a natural approach. For each $z \in R^m$, the dual objective function can be computed in a pointwise way. Indeed, solving the maximization problem involved in $\phi(z)$

$$\phi(z) = \sup_{u\in S}\{-J(u) - \langle \mathcal{A}u, z\rangle\} = \sup_{u(t)\in S(t)\ \forall t\in[0,\mathcal{T}]} -\int_0^{\mathcal{T}} F(t, u(t)) - \langle A(t)u(t), z\rangle dt$$

is equivalent to solving independently for each value of $t \in [0,\mathcal{T}]$, a subproblem over $S(t) \subset R^m$: $\max_{v\in S(t)}[-F(t, v) - \langle v, A(t)^T z\rangle]$, which is assumed to be easy to solve or even computable in closed form. The same separability property holds after smoothing, since minimizing the smoothed dual objective function $\sup_{u\in S}\{-J(u) - \langle \mathcal{A}u, z\rangle - \frac{\mu}{2}\|u\|^2_U\}$ is equivalent to solving pointwise subproblems $\max_{v\in S(t)}\{-F(t, v) - \langle v, A(t)^T z\rangle - \frac{\mu}{2}\|v\|^2_{R^m}\}$.

3.4 Double Smoothing Technique

We propose to solve the dual problem (3.9) using a new primal-dual smoothing technique. Note that, as shown above, its objective function is in general not

differentiable and not strongly convex. However, we can ensure these properties by a double primal-dual regularization of θ. The goal of the first regularization is to obtain an objective function with Lipschitz-continuous gradient, for which we can apply much more efficient algorithms of smooth convex optimization. The goal of the second regularization is to obtain a strongly convex dual objective. This property is necessary to allow us to reconstruct efficiently a nearly feasible and optimal primal solution from a nearly optimal dual solutions.

3.4.1 First Smoothing

Let us start by ensuring smoothness of the dual function. Dual objective $\theta(z)$ is a sum of two functions, both of which can be nonsmooth. Their nonsmoothness comes from the fact that the optimization problems defining $\sigma_T(z)$ and $\phi(z)$ can have multiple optimal solutions at a given point z. A natural way to obtain a smooth approximation of θ is to modify these optimization subproblems in order to ensure the uniqueness of optimal solutions for each $z \in V$. For any $\rho > 0$, we can approximate $\sigma_T(z) = \sup_{x \in T} \langle x, z \rangle$ by a modified function

$$\sigma_{\rho,T}(z) = \sup_{x \in T} \{\langle x, z \rangle - \frac{\rho}{2} \|x\|_{V^*}^2\}. \tag{3.11}$$

In the same way, for any $\mu > 0$, we modify function $\phi(z)$ as follows

$$\phi_\mu(z) = \sup_{u \in S} \{-J(u) - \langle \mathcal{A}u, z \rangle - \frac{\mu}{2} \|u\|_U^2\}. \tag{3.12}$$

The following result can be seen as an easy generalization of Theorem 1 in [59]:

Theorem 3.2. *Let H_1, H_2 be two Hilbert spaces. Assume that the linear operator $A : H_1 \to H_2^*$ is bounded and that the function $G : H_1 \to R$ is closed and strongly convex with parameter κ. Let $Q \subset \mathrm{dom}\,(G)$ be a closed, convex set. Then the function*

$$F(z) = \sup_{u \in Q} \{-G(u) - \langle \mathcal{A}u, z \rangle\} \tag{3.13}$$

is smooth with Lipschitz-continuous gradient $\nabla F(z) = -\mathcal{A}u_z$ where u_z is the unique optimal solution of the optimization problem defining $F(z)$. The Lipschitz constant of the gradient is equal to $\frac{\|A\|^2}{\kappa}$ where $\|A\| = \sup\{\langle \mathcal{A}u, z \rangle : u \in H_1, \|u\|_{H_1} = 1, z \in H_2, \|z\|_{H_2} = 1\}$.

Choosing now $H_1 = U$, $Q = S$, $H_2 = V$, $A = \mathcal{A}$ and $G(u) = J(u) + \frac{\mu}{2} \|u\|_U^2$, we conclude that ϕ_μ is smooth and convex with a Lipschitz continuous gradient equal to $\nabla \phi_\mu(z) = -\mathcal{A}u_{\mu,z}$, where $u_{\mu,z}$ denotes the unique optimal solution of the problem (3.12). Its Lipschitz constant is given by $L(\phi_\mu) = \frac{\|\mathcal{A}\|^2}{\mu}$. Similarly,

if we choose $H_1 = V^*$, $Q = T$, $H_2 = V$, $A = I : V^* \to V^*$ and $G(x) = \frac{\rho}{2}\|x\|^2_{V^*}$, Theorem 5.1 shows that $\sigma_{\rho,T}$ is smooth and convex with Lipschitz-continuous gradient $\nabla\sigma_{\rho,T}(z) = x_{\rho,z}$, where $x_{\rho,z}$ denotes the unique optimal solution of the problem (3.11). The Lipschitz constant of this gradient is given by $L(\sigma_{\rho,T}) = \frac{1}{\rho}$.

Remark 3.2. When the function $J(u)$ is strongly convex with parameter κ, we do not need to apply the first smoothing to $\phi(z)$, which is in this case already smooth with a Lipschitz-continuous gradient with constant $\frac{\|A\|^2}{\kappa}$.

If we denote $D_T = \max\{\frac{1}{2}\|x\|^2_{V^*} : x \in T\}$ and $D_S = \max\{\frac{1}{2}\|u\|^2_U : u \in S\}$, it is easy to check that $\sigma_{\rho,T}(z) \leq \sigma_T(z) \leq \sigma_{\rho,T}(z) + \rho D_T$ for all $z \in V$ and $\phi_\mu(z) \leq \phi(z) \leq \phi_\mu(z) + \mu D_S$ for all $z \in V$. Therefore, if we define the regularized function

$$\theta_{\rho,\mu}(z) = \sigma_{\rho,T}(z) + \phi_\mu(z)$$

we have $\theta_{\rho,\mu} \in F^{1,1}_{L(\rho,\mu)}(V)$ with $L(\rho,\mu) := \frac{1}{\rho} + \frac{\|\mathcal{A}\|^2}{\mu}$ and

$$\theta_{\rho,\mu}(z) \leq \theta(z) \leq \theta_{\rho,\mu}(z) + \mu D_S + \rho D_T \quad \forall z \in V. \tag{3.14}$$

Applying a fast gradient method to the function $\theta_{\rho,\mu}$ will generate a point $z_\epsilon \in V$ such that $\theta(z_\epsilon) - \theta^* \leq \epsilon$ in $O\left(\frac{1}{\epsilon}\right)$ iterations (see [59]). However, our aim is not only to solve the dual problem efficiently but also to generate a nearly optimal and nearly feasible solution for the primal problem. We will see that a single smoothing is not enough in order to achieve this goal.

Let us first show how it is possible to reconstruct a good primal solution from a dual iterate. Let $z \in V$, we have:

$$\begin{aligned}\theta_{\rho,\mu}(z) &= \sigma_{\rho,T}(z) + \phi_\mu(z)\\ &= \langle x_{\rho,z}, z\rangle - \frac{\rho}{2}\|x_{\rho,z}\|^2_{V^*} - J(u_{\mu,z}) - \langle \mathcal{A}u_{\mu,z}, z\rangle - \frac{\mu}{2}\|u_{\mu,z}\|^2_U.\end{aligned}$$

Let us find the conditions that z must satisfy in order to guarantee that $u_{\mu,z}$ is nearly optimal and nearly feasible for the primal problem. We have:

$$\begin{aligned}J(u_{\mu,z}) - D^* &= \langle x_{\rho,z}, z\rangle - \frac{\rho}{2}\|x_{\rho,z}\|^2_{V^*} - \theta_{\rho,\mu}(z) - \langle \mathcal{A}u_{\mu,z}, z\rangle - \frac{\mu}{2}\|u_{\mu,z}\|^2_U + \theta^*\\ &= \langle \nabla\theta_{\rho,\mu}(z), z\rangle + (\theta^* - \theta_{\rho,\mu}(z)) - \frac{\rho}{2}\|x_{\rho,z}\|^2_{V^*} - \frac{\mu}{2}\|u_{\mu,z}\|^2_U.\end{aligned}$$

Therefore, since $J(u_{\mu,z}) \geq P^* \geq D^*$, we have $J(u_{\mu,z}) - D^* \geq J(u_{\mu,z}) - P^* \geq 0$ and

$$J(u_{\mu,z}) - P^* \leq |\langle \nabla\theta_{\rho,\mu}(z), z\rangle| + |\theta_{\rho,\mu}(z) - \theta^*| + \rho D_T + \mu D_S$$

and, as we also have that $|\theta_{\rho,\mu}(z) - \theta^*| \leq |\theta(z) - \theta^*| + \mu D_S + \rho D_T$, we conclude that

$$J(u_{\mu,z}) \leq P^* + |\langle \nabla\theta_{\rho,\mu}(z), z\rangle| + (\theta(z) - \theta^*) + 2\rho D_T + 2\mu D_S.$$

If we apply the fast gradient method to the function $\theta_{\rho,\mu}$ with constants ρ and μ chosen to be of order $O(\frac{1}{k})$, we already know from [59] that the kth iterate z_k generated by this algorithm satisfies $\theta(z_k) - \theta^* \leq O(\frac{1}{k})$. However, unfortunately, the norm of the gradient of the smoothed function $\|\nabla\theta_{\rho,\mu}(z_k)\|_V$ does not decrease at the same rate, which negatively affects the above estimate of the rate of convergence of $J(u_{\mu,z})$. Indeed, as $\theta_{\rho,\mu} \in F^{1,1}_{L(\rho,\mu)}(V)$, we have (see Theorem 2.1.5 in [58])

$$\|\nabla\theta_{\rho,\mu}(z_k)\|^2_{V^*} \leq 2L(\rho,\mu)(\theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu}).$$

As the fast gradient method is applied to the function $\theta_{\rho,\mu}$, we also have ([59])

$$\theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu} \leq \frac{4L(\rho,\mu)\,\|z_0 - z^*_S\|}{(k+1)(k+2)},$$

where z^*_S denotes any optimal solution of the smoothed dual problem $\min_{z\in V}\theta_{\rho,\mu}(z)$, and we see that choosing $L(\rho,\mu)$ of order $O(k)$ guarantees convergence of the smoothed objective of order $O(\frac{1}{k})$. Therefore

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \frac{2\sqrt{2}L(\rho,\mu)\sqrt{\|z_0 - z^*_S\|}}{\sqrt{(k+1)(k+2)}}$$

and, due to the fact that $L(\rho,\mu)$ is of order $O(k)$, we cannot guarantee that the norm of the gradient $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}$ is decreasing with respect to k.

In principle, this can be remedied with a minor modification of the scheme. It is indeed possible to obtain in $2k$ iterations a point $\tilde{z}$ such that $\|\nabla\theta_{\rho,\mu}(\tilde{z})\|_{V^*}$ is of order $O(\frac{1}{\sqrt{k}})$. Indeed, after k steps of the fast gradient method have been computed, we can apply k additional steps of the classical gradient method with constant stepsize $\frac{1}{L(\rho,\mu)}$, i.e.

$$z_{k+i} = z_{k+i-1} - \frac{1}{L(\rho,\mu)}\nabla\theta_{\rho,\mu}(z_{k+i-1}) \quad i = 1,\ldots,k\,.$$

We have then (see Theorem 2.1.14 in [58])

$$\frac{1}{2L(\rho,\mu)}\,\|\nabla\theta_{\rho,\mu}(z_{k+i-1})\|^2_{V^*} \leq \theta_{\rho,\mu}(z_{k+i-1}) - \theta_{\rho,\mu}(z_{k+i}) \quad \forall i = 1,\ldots,k.$$

and summing these inequalities gives

$$\begin{aligned}\sum_{i=1}^{k} \frac{1}{2L(\rho,\mu)} \left\|\nabla\theta_{\rho,\mu}(z_{k+i})\right\|_{V^*}^2 &\leq \theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z_{2k}) \\ &\leq \theta_{\rho,\mu}(z_k) - \theta^*_{\rho,\mu} \\ &\leq \frac{4L(\rho,\mu)\left\|z_0 - z_S^*\right\|_V^2}{(k+1)(k+2)}.\end{aligned}$$

If we denote by $\tilde{z}$ the iterate with the smallest norm of the gradient, we conclude that

$$\left\|\nabla\theta_{\rho,\mu}(\tilde{z})\right\|_{V^*}^2 = \min_{i=1,\ldots,k} \left\|\nabla\theta_{\rho,\mu}(z_{k+i})\right\|_{V^*}^2 \leq \frac{8L^2(\rho,\mu)\left\|z_0 - z_S^*\right\|_V^2}{k(k+1)(k+2)}.$$

In conclusion, after $2k$ iterates, we are able mixing fast and classical gradient method, to obtain a point $\tilde{z}$ such that $\theta_{\rho,\mu}(\tilde{z}) - \theta^*_{\rho,\mu} \leq O(\frac{1}{k})$ and $\left\|\nabla\theta_{\rho,\mu}(\tilde{z})\right\|_{V^*} = O(\frac{1}{\sqrt{k}})$. However, this convergence is very slow, as it implies we need at least $k = O(\frac{1}{\epsilon^2})$ iterations in order to have a primal solution $u_{\mu,z_k} \in S$ with accuracy ϵ (i.e. $J(u_{\mu,z_k}) \leq P^* + \epsilon$). This is not better than the result of the classical subgradient approach.

Furthermore, we must also examine the feasibility of the reconstructed primal iterate $u_{\mu,\tilde{z}}$. The norm of $\left\|\nabla\theta_{\rho,\mu}(\tilde{z})\right\|_{V^*}$ also provides an upper bound for the non admissibility measure of $u_{\mu,\tilde{z}}$. Indeed, denoting $d(\cdot,T)$ the distance to set T, we have

$$d(\mathcal{A}u_{\mu,\tilde{z}},T) \leq \left\|\mathcal{A}u_{\mu,\tilde{z}} - x_{\rho,\tilde{z}}\right\|_V = \left\|\nabla\theta_{\rho,\mu}(\tilde{z})\right\|_V .$$

In conclusion, we have shown why requiring a good convergence rate for $\theta(z_k) - \theta^*$ alone is not sufficient to obtain a nearly feasible and optimal solution for the primal problem. The same good rate must also be ensured for the norm of the gradient $\left\|\nabla\theta_{\rho,\mu}(z_k)\right\|_V$. We now show how applying a second smoothing to the dual objective function, making it also strongly convex, will achieve this goal.

3.4.2 Second Smoothing

In order to obtain a strongly convex dual objective function, we simply add the strongly convex function $\frac{\kappa}{2}\left\|z\right\|_V^2$ to the function $\theta_{\rho,\mu}$. This gives us a new dual objective function:

$$\theta_{\rho,\mu,\kappa}(z) = \sigma_{\rho,T}(z) + \phi_\mu(z) + \frac{\kappa}{2}\left\|z\right\|_V^2 ,$$

which is strongly convex with parameter κ. If we denote by $B = B^* : V \to V^*$ (with $B \succ 0$) the duality map between V and its dual space, i.e. $\langle Bz, \overline{z}\rangle = (z|\overline{z})_V \ \forall z, \overline{z} \in V$, we have $\nabla\theta_{\rho,\mu,\kappa}(z) = x_{\rho,z} - \mathcal{A}u_{\mu,z} + \kappa Bz$. This gradient is Lipschitz-continuous with constant $L(\rho,\mu,\kappa) := \frac{1}{\rho} + \frac{\|\mathcal{A}\|^2}{\mu} + \kappa$, hence this function belongs to $S^{1,1}_{\kappa,L(\rho,\mu,\kappa)}(V)$. Denote by $\theta^*_{\rho,\mu,\kappa}$ the optimal value of problem $\min_{z\in S}\theta_{\rho,\mu,\kappa}(z)$. Applying a fast gradient method for the class $S^{1,1}_{\kappa,L(\rho,\mu),\kappa}$ to function $\theta_{\rho,\mu,\kappa}$, we generate a sequence z_k satisfying

$$\theta_{\rho,\mu,\kappa}(z_k) - \theta^*_{\rho,\mu,\kappa} \overset{(3.4)}{\leq} \exp\left(-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right) 2(\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa})$$

and

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|^2_{V^*} \overset{(3.5)}{\leq} 4L(\rho,\mu,\kappa)(\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa}) \exp\left(-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)$$

i.e.

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} \leq 2\sqrt{L(\rho,\mu,\kappa)}\sqrt{\theta_{\rho,\mu,\kappa}(z_0) - \theta^*_{\rho,\mu,\kappa}} \exp\left(-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}\right)$$

and we have the same rate of convergence for $\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*}$ as for $\theta_{\rho,\mu,\kappa}(z_k) - \theta^*_{\rho,\mu,\kappa}$. This property is crucial in order to obtain a nearly feasible and optimal primal solution in $O\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right)\right)$ iterations (instead of $O\left(\frac{1}{\epsilon^2}\right)$ with a simple smoothing).

3.5 Strong duality and norm of dual optimal solutions

Before we apply the fast gradient method to the doubly smoothed dual function, we study in this section under which condition strong duality, i.e $P^* = D^*$, holds and how it is possible to bound the size of the optimal solution set of the dual problem (3.9) . Such bound will play a role in the convergence analysis of our scheme, as we will see in the following section.

Theorem 3.3. *If there exists $r > 0$ such that*

$$B(0,r) \subset Q := \{x - \mathcal{A}u : u \in S, x \in T\} \subset V^* \tag{3.15}$$

then

- ◇ *there is no duality gap, i.e. $P^* = D^*$, and*

- ◇ *the optimal solution set of the dual problem (3.9) is a nonempty, bounded, closed and convex set in V.*

Furthermore, if we define $\Delta(J') := \max_{u,v\in S} J'(v, u-v)$, where $J'(v, u-v)$ is the directional derivative of J at point v in direction $u-v$, we have the following upper-bound for the norm of the dual optimal solutions

$$\|z^*\|_V \leq \frac{\Delta(J')}{r}.$$

Proof. Applying Theorem 2.165 in [11] to the primal problem $\min_{u\in U}\{f(u) = J(u) + I_S(u) : G(u) = Au \in T\}$ (where I_S denotes the indicator function of S), we conclude, using our assumptions on J, A, S and T and the regularity condition (3.15), that there is no duality gap between this problem and its Lagrangian dual (3.9). Furthermore, as the primal optimal value P^* is assumed to be finite, we have that the optimal solution set of the dual problem is a non-empty, bounded, closed and convex set in V.

It remains to obtain the bound $\|z^*\|_V \leq \frac{\Delta(J')}{r}$. As the subproblems $\sup_{x\in T}\langle x, z\rangle$ and $\sup_{u\in S}\{-J(u) - \langle \mathcal{A}u, z\rangle\}$ are solvable for all $z \in V$,

$$\partial\theta(z) = \{x_z - \mathcal{A}u_z \in V^* : x_z \in T, \langle x_z, z\rangle = \sigma_T(z), u_z \in S, -J(u_z) - \langle \mathcal{A}u_z, z\rangle = \phi(z)\}$$

is non-empty for all $z \in V$. Let $g_z = x_z - \mathcal{A}u_z$ be any element in $\partial\theta(z)$. By the optimality condition of the problems defining $\sigma_T(z)$ and $\phi(z)$ we write

$$\langle x - x_z, z\rangle \leq 0 \quad \forall x \in T \qquad -J'(u_z, u-u_z) - \langle \mathcal{A}(u-u_z), z\rangle \leq 0 \quad \forall u \in S.$$

This can be rewritten

$$\langle x_z, z\rangle \geq \langle x, z\rangle \quad \forall x \in T \quad -\langle \mathcal{A}u_z, z\rangle \geq -J'(u_z, u-u_z) + \langle -\mathcal{A}u, z\rangle \quad \forall u \in S$$

and therefore

$$\begin{aligned} \langle g_z, z\rangle = \langle x_z - \mathcal{A}u_z, z\rangle &\geq \langle x, z\rangle - J'(u_z, u-u_z) + \langle -\mathcal{A}u, z\rangle \\ &= \langle x - \mathcal{A}u, z\rangle - J'(u_z, u-u_z). \end{aligned}$$

We obtain:

$$\langle x - Au, z\rangle \leq J'(u_z, u-u_z) + \langle g_z, z\rangle \quad \forall u \in S, \quad \forall x \in T$$

and therefore

$$\max_{u\in S, x\in T}\langle x - \mathcal{A}u, z\rangle \leq \max_{u\in S} J'(u_z, u-u_z) + \langle g_z, z\rangle \leq \max_{u,v\in S} J'(v, u-v) + \langle g_z, z\rangle$$

As $B(0, r) \subset \{x - \mathcal{A}u : u \in S, x \in T\} \subset V^*$, we have $\max_{u\in S, x\in T}\langle x - \mathcal{A}u, z\rangle \geq r\|z\|_V$ and therefore

$$r\|z\|_V \leq \max_{u,v\in S} J'(v, u-v) + \langle g_z, z\rangle \quad \forall z \in V. \tag{3.16}$$

Consider now an optimal solution z^* of the dual problem. By the optimality condition of this problem, we have: $0 \in \partial\theta(z^*)$ and therefore $\|z^*\|_V \leq \frac{\Delta(J')}{r}$. □

Remark 3.3. As J is convex, we have $J(u) \geq J(v) + J'(v, u-v)\ \forall u, v \in S$ and

$$\max_{u,v\in S}(J(u) - J(v)) = \max_{u\in S} J(u) - \min_{v\in S} J(v) \geq \max_{u,v\in S} J'(v, u-v).$$

The condition $\max_{u,v\in S} J'(v, u-v) < +\infty$ is therefore satisfied as soon as J has bounded variation on S, i.e. $\max_{u\in S} J(u) - \min_{v\in S} J(v) < +\infty$. However, our condition is strictly weaker, as the following example shows. Consider the function

$$J(u) = \int_0^1 \ln(1 - u^2(t))dt$$

defined on $S = \{u \in L^2([0,1]) : -1 \leq u(t) \leq 1 \quad a.e \quad in \quad [0,1]\}$. Clearly $\max_{u\in S} J(u) - \min_{v\in S} J(v) = +\infty$, and this function can be seen as as barrier function for S. However we have:

$$J'(v, u-v) = \int_0^1 \frac{2(u(t) - v(t))v(t)}{1 - v^2(t)}dt$$

and $\Delta(J') = \max_{u,v\in S'} J'(v, u-v) \leq 2$.

Remark 3.4. If the primal problem is feasible, it is clear that there exist $\overline{u} \in S$ and $\overline{x} \in T$ such that $\mathcal{A}\overline{u} = \overline{x}$, i.e $0 \in Q$. In order to have $B(0,r) \subset Q$ with $r > 0$, one of the following extra assumptions is enough:

- ◇ **The set T has a non-empty interior and there exists $\overline{u} \in S$ such that $\mathcal{A}\overline{u} = \overline{x} \in \text{int } T$ (generalized Slater condition).**
  As $\overline{x} \in \text{int } T$, there exists $r > 0$ such that: $\overline{x} + B(0,r) \subset T$ and therefore $B(0,r) = \overline{x} - \mathcal{A}\overline{u} + B(0,r) \subset Q$.

- ◇ **The set S has a non-empty interior, $\mathcal{A} : U \to V^*$ is surjective and there exists $\overline{u} \in \text{int } S$ such that $\mathcal{A}\overline{u} = \overline{x} \in T$.**
  Indeed in this case, as $\overline{u} \in \text{int } S$, there exists $\tilde{r} > 0$ such that $B(\overline{u}, \tilde{r}) \subset S$. By the Banach-Schauder theorem, the image of any open subset of U by $\mathcal{A}$ is an open subset in V^*. Therefore, there exists $r > 0$ such that $\mathcal{A}\overline{u} + B(0,r) \subset \mathcal{A}(B(\overline{u}, \tilde{r})) \subset \mathcal{A}(S)$. We conclude that $\overline{x} - \mathcal{A}\overline{u} + B(0,r) = B(0,r) \subset Q$.

3.6 Solving the primal-dual problem in $O(\frac{1}{\epsilon}\ln\left(\frac{1}{\epsilon}\right))$ iterations

Denote by z^*_{DS} the unique optimal solution of the doubly smoothed dual problem

$$\min_{z\in V}\theta_{\rho,\mu,\kappa}(z), \tag{3.17}$$

and by z^* one of the optimal solutions of the original dual problem (3.9). We assume that the upper bound

$$\|z^*\|_V \quad\leq\quad D_D \tag{3.18}$$

is available. As we have just shown, this can be ensured under very natural assumptions on S, T and J using Theorem 3.3.
If we apply the simple fast gradient method presented in subsection 3.1 to the doubly smoothed dual problem with starting point $z_0 = 0$, we obtain a sequence $\{z_k\}$ verifying

$$\begin{aligned}
\theta_{\rho,\mu,\kappa}(z_k) - \theta_{\rho,\mu,\kappa}(z^*_{DS}) \;&\leq\; 2\left(\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})\right) e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}},\\
\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|^2_{V^*} \;&\leq\; 4L(\rho,\mu,\kappa)(\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS}))\; e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}},\\
\|z_k - z^*_{DS}\|^2_{V^*} \;&\leq\; \min\left\{\|z^*_{DS}\|^2_{V^*}, \tfrac{4}{\kappa}\left(\theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})\right) e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}\right\}.
\end{aligned} \tag{3.19}$$

3.6.1 Convergence of $\theta(z_k)$ to θ^*

Since $\theta_{\rho,\mu,\kappa}(0) = \theta_{\rho,\mu}(0)$ and $\theta_{\rho,\mu,\kappa}(z^*_{DS}) = \theta_{\rho,\mu}(z^*_{DS}) + \frac{\kappa}{2}\|z^*_{DS}\|^2_V$, we have

$$\begin{aligned}
\tfrac{\kappa}{2}\|z^*_{DS}\|^2_V \quad&\leq\quad \theta_{\rho,\mu,\kappa}(0) - \theta_{\rho,\mu,\kappa}(z^*_{DS})\\
&=\quad \theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}) - \tfrac{\kappa}{2}\|z^*_{DS}\|^2_V,\\
\|z_k - z^*_{DS}\|^2_V \quad&\overset{(3.4)}{\leq}\quad \tfrac{4}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}.
\end{aligned} \tag{3.20}$$

Note that

$$\begin{aligned}
\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \quad&\overset{(3.4)}{\leq}\quad 2(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}\\
&\qquad +\frac{\kappa}{2}(\|z^*_{DS}\|^2_V - \|z_k\|^2_V).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|z^*_{DS}\|^2_V - \|z_k\|^2_V \quad&\leq\quad \|z^*_{DS} - z_k\|_V(\|z^*_{DS}\|_V + \|z_k\|_V)\\
&\leq\quad \|z^*_{DS} - z_k\|_V(2\|z^*_{DS}\|_V + \|z_k - z^*_{DS}\|_V)\\
&\overset{(3.6)}{\leq}\quad 3\|z^*_{DS} - z_k\|_V\cdot\|z^*_{DS}\|_V\\
&\overset{(3.20)}{\leq}\quad 3\cdot\|z^*_{DS}\|_V\sqrt{\tfrac{4}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))}e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}\\
&\overset{(3.20)}{\leq}\quad \tfrac{6}{\kappa}(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}},
\end{aligned}$$

and therefore

$$\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \quad \leq \quad 5(\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}.$$

We also have $\theta_{\rho,\mu}(0) \leq \theta(0)$ and

$$\theta_{\rho,\mu}(z^*_{DS}) \quad \geq \quad \theta(z^*_{DS}) - \rho D_T - \mu D_S \ \geq \ \theta(z^*) - \rho D_T - \mu D_S.$$

Therefore,

$$\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}) \quad \leq \quad \theta(0) - \theta(z^*) + \rho D_T + \mu D_S. \tag{3.21}$$

Finally, since $\theta_{\rho,\mu}(z^*_{DS}) + \frac{\kappa}{2}\|z^*_{DS}\|^2_V \leq \theta_{\rho,\mu}(z^*) + \frac{\kappa}{2}\|z^*\|^2_V$, we have

$$\theta_{\rho,\mu}(z^*_{DS}) \quad \leq \quad \theta_{\rho,\mu}(z^*) + \tfrac{\kappa}{2}\|z^*\|^2_V \quad \overset{(3.14)}{\leq} \quad \theta(z^*) + \tfrac{\kappa}{2}\|z^*\|^2_V,$$

and therefore

$$\theta_{\rho,\mu}(z_k) - \theta_{\rho,\mu}(z^*_{DS}) \quad \overset{(3.14)}{\geq} \quad \theta(z_k) - \mu D_S - \rho D_T - \theta(z^*) - \tfrac{\kappa}{2}\|z^*\|^2_V.$$

In conclusion, we have

$$\begin{aligned} \theta(z_k) - \theta(z^*) \leq \mu D_S + \rho D_T + \tfrac{\kappa}{2} D_D^2 \\ +5\left(\theta(0) - \theta(z^*) + \rho D_T + \mu D_S\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \end{aligned} \tag{3.22}$$

Now it is clear how to choose the smoothing parameters. Let us fix some $\epsilon > 0$. In the upper bound for the residual $\theta(z_k) - \theta(z^*)$, we have four terms. In order to ensure accuracy $\theta(z_k) - \theta(z^*) \leq \epsilon$, we force all of these terms to be less or equal than $\frac{\epsilon}{4}$. This leads to the following values:

$$\mu = \mu(\epsilon) = \tfrac{\epsilon}{4D_S}, \quad \rho = \rho(\epsilon) = \tfrac{\epsilon}{4D_T}, \quad \kappa = \kappa(\epsilon) = \tfrac{\epsilon}{2D_D^2}. \tag{3.23}$$

Under this choice we get

$$\theta(z_k) - \theta(z^*) \quad \leq \quad \tfrac{3\epsilon}{4} + 5\left(\theta(0) - \theta(z^*) + \tfrac{\epsilon}{2}\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \tag{3.24}$$

The last term in the estimate (3.24) defines the number of iterations needed for reaching the accuracy ϵ. Clearly, we ensure
$5\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right) e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \leq \frac{\epsilon}{4}$ by taking

$$k \quad \geq \quad 2\sqrt{\tfrac{L(\rho,\mu,\kappa)}{\kappa}} \ln \tfrac{20\left(\theta(0) - \theta(z^*) + \frac{\epsilon}{2}\right)}{\epsilon}. \tag{3.25}$$

It remains to note that

$$\tfrac{L(\rho,\mu,\kappa)}{\kappa} \quad = \quad 1 + \tfrac{1}{\rho\kappa} + \tfrac{1}{\mu\kappa}\|\mathcal{A}\|^2 \overset{(3.23)}{=} 1 + \tfrac{8}{\epsilon^2}\left[D_T + D_S\|\mathcal{A}\|^2\right] D_D^2. \tag{3.26}$$

Thus, we need at most $k = O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ iterations.

3.6.2 Convergence of $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*}$

In our approach, we want to be able to reconstruct a nearly optimal and feasible primal solution efficiently. In Section 3.4, we have seen that the accuracy of this primal solution depends not only on the rate of convergence for the dual objective function, but also on the rate of convergence of the norm of its gradient. We provide now an upper bound on the number of iterations needed to reduce this norm below a certain level. We start with

$$\begin{array}{rcl} \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} & \leq & \|\nabla\theta_{\rho,\mu,\kappa}(z_k) - \kappa B z_k\|_{V^*} \\ & \leq & \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|Bz_k\|_{V^*} \\ & = & \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|z_k\|_V \\ & \overset{(3.19)}{\leq} & \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\left\|z^*_{DS}\right\|_V. \end{array}$$

Note that

$$\begin{array}{rcl} \frac{1}{4L(\rho,\mu,\kappa)}\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|^2_{V^*} & \overset{(3.19),(3.20)}{\leq} & (\theta_{\rho,\mu}(0) - \theta_{\rho,\mu}(z^*_{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \\ & \overset{(3.21)}{\leq} & (\theta(0) - \theta(z^*) + \mu D_S + \rho D_T)e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} \\ & \overset{(3.23)}{=} & (\theta(0) - \theta(z^*) + \frac{\epsilon}{2})e^{-k\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}}. \end{array}$$

At the same time,

$$\begin{array}{rcl} \theta(z^*) + \frac{\kappa}{2}\left\|z^*\right\|^2_V & \overset{(3.14)}{\geq} & \theta_{\rho,\mu}(z^*) + \frac{\kappa}{2}\left\|z^*\right\|^2_V \;\geq\; \theta_{\rho,\mu}(z^*_{DS}) + \frac{\kappa}{2}\left\|z^*_{DS}\right\|^2_V \\ & \overset{(3.14)}{\geq} & \theta(z^*_{DS}) - \mu D_S - \rho D_T + \frac{\kappa}{2}\left\|z^*_{DS}\right\|^2_V \\ & \geq & \theta(z^*) - \mu D_S - \rho D_T + \frac{\kappa}{2}\left\|z^*_{DS}\right\|^2_2. \end{array}$$

Hence,

$$\|z^*_{DS}\|_V \;\leq\; \sqrt{\|z^*\|^2_2 + \tfrac{2\mu}{\kappa}D_S + \tfrac{2\rho}{\kappa}D_T} \overset{(3.23)}{\leq} \kappa^{-1/2}\sqrt{\tfrac{3\epsilon}{2}} \overset{(3.23)}{=} \sqrt{3}D_D, \tag{3.27}$$

and we obtain:

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \sqrt{4L(\rho,\mu,\kappa)(\theta(0) - \theta(z^*) + \tfrac{\epsilon}{2})}e^{-\frac{k}{2}\sqrt{\frac{\kappa}{L(\rho,\mu,\kappa)}}} + 2\sqrt{3}\kappa D_D.$$

Taking into account (3.23), we can see that in $k(\epsilon) = O(\frac{1}{\epsilon}\ln\frac{1}{\epsilon})$ iterations, we can ensure

$$\theta(z_k) - \theta(z^*) \leq \epsilon, \quad \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \tfrac{2\epsilon}{D_D}. \tag{3.28}$$

3.6.3 Constructing an approximate primal solution

The possibility to construct a nearly feasible and optimal primal solution from a nearly optimal dual solution with the same level of accuracy is the crucial advantage of the double smoothing technique compared to a simple smoothing. In this section, given a target accuracy $\epsilon > 0$, we will see how to obtain from the dual iterate $z_{k(\epsilon)}$ an approximate primal solution $\widehat{u}_{k(\epsilon)} \in S$ such that

$$\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| \quad \leq \quad 2\left(1+2\sqrt{3}\right)\cdot\epsilon, \tag{3.29}$$

$$d(\mathcal{A}\widehat{u}_{k(\epsilon)}, T) \leq \frac{2\epsilon}{D_D} \tag{3.30}$$

Since $D^* \leq P^*$, inequality (3.29) implies $J(\widehat{u}_{k(\epsilon)}) \leq P^* + 2\left(1+2\sqrt{3}\right)\cdot\epsilon$, and $\hat{u}_{k(\epsilon)}$ satisfying (3.29), (3.30) can be seen as a nearly optimal and feasible primal solution with accuracy proportional to ϵ.

Consider $\widehat{u}_{k(\epsilon)} = u_{\mu(\epsilon),z_{k(\epsilon)}}$, the unique optimal solution of the optimization problem defining $\phi_{\mu(\epsilon)}(z_{k(\epsilon)})$. We have

$$\begin{array}{rcl}\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) & = & \sigma_{\rho(\epsilon),T}(z_{k(\epsilon)}) + \phi_{\mu}(z_{k(\epsilon)}) \\ & = & \langle x_{\rho(\epsilon),z_{k(\epsilon)}}, z_{k(\epsilon)}\rangle - \frac{\rho(\epsilon)}{2}\left\|x_{\rho(\epsilon),z_{k(\epsilon)}}\right\|^2_{V^*} - J(\widehat{u}_{k(\epsilon)}) \\ & & -\langle \mathcal{A}\widehat{u}_{k(\epsilon)}, z_{k(\epsilon)}\rangle - \frac{\mu(\epsilon)}{2}\left\|\widehat{u}_{k(\epsilon)}\right\|^2_U.\end{array}$$

Therefore,

$$\begin{array}{rcl}J(\widehat{u}_{k(\epsilon)}) - D^* & = & \langle x_{\rho(\epsilon),z_{k(\epsilon)}} - \mathcal{A}\widehat{u}_{k(\epsilon)}, z_{k(\epsilon)}\rangle - \frac{\rho(\epsilon)}{2}\left\|x_{\rho(\epsilon),z_{k(\epsilon)}}\right\|^2_{V^*} \\ & & -\frac{\mu(\epsilon)}{2}\left\|\widehat{u}_{k(\epsilon)}\right\|^2_U - \theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) + \theta(z^*).\end{array} \tag{3.31}$$

Since $\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*) \leq \theta(z_{k(\epsilon)}) - \theta(z^*) \leq \epsilon$, and

$$\begin{array}{rcl}\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*) & \overset{(3.14)}{\geq} & \theta(z_{k(\epsilon)}) - \mu(\epsilon)D_S - \rho(\epsilon)D_T - \theta(z^*) \\ & \overset{(3.23)}{=} & \theta(z_{k(\epsilon)}) - \theta(z^*) - \frac{1}{2}\epsilon \;\geq\; -\frac{1}{2}\epsilon,\end{array}$$

we have $|\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)}) - \theta(z^*)| \leq \epsilon$. Therefore,

$$\begin{array}{rcl}\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| & \leq & \left\|x_{\rho(\epsilon),z_{k(\epsilon)}} - \mathcal{A}\widehat{u}_{k(\epsilon)}\right\|_{V^*}\left\|z_{k(\epsilon)}\right\|_V \\ & & +\rho(\epsilon)\hat{D}_T + \mu(\epsilon)D_S + \epsilon \\ & \overset{(3.23)}{\leq} & \left\|\nabla\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)})\right\|_{V^*}\left\|z_{k(\epsilon)}\right\|_V + 2\epsilon \\ & \overset{(3.28)}{\leq} & \frac{2\epsilon}{D_D}\left\|z_{k(\epsilon)}\right\|_V + 2\epsilon,\end{array}$$

establishing (3.29). On the other hand,

$$\left\|z_{k(\epsilon)}\right\|_V \quad \leq \quad \left\|z_{k(\epsilon)} - z^*_{DS}\right\|_V + \left\|z^*_{DS}\right\|_V \overset{(3.19)}{\leq} 2\left\|z^*_{DS}\right\|_V \overset{(3.27)}{\leq} 2\sqrt{3}D_D$$

and we obtain $\left|J(\widehat{u}_{k(\epsilon)}) - D^*\right| \quad \leq \quad 2\left(1+2\sqrt{3}\right)\cdot\epsilon$. Finally, we have $\widehat{u}_{k(\epsilon)} \in S$ and

$$\left\|\mathcal{A}\widehat{u}_{k(\epsilon)} - x_{\rho(\epsilon),z_{k(\epsilon)}}\right\|^2_{V^*} = \left\|\nabla\theta_{\rho(\epsilon),\mu(\epsilon)}(z_{k(\epsilon)})\right\|^2_{V^*} \overset{(3.28)}{\leq} \left(\tfrac{2\epsilon}{\overline{D}_D}\right)^2,$$

where $x_{\rho(\epsilon),z_{k(\epsilon)}} \in T$, which proves (3.30).

3.7 Practical Implementation of the Double Smoothing Technique

The double smoothing technique presented in the previous section relies on a choice of the smoothing parameter κ that requires knowledge of the size of the dual optimal set, or at least an upper-bound D_D on this quantity (given for example by Theorem 3.3).

Even when we have no initial estimate of the size of $\|z^*\|_V$, we can still apply a simple modification of the double smoothing technique. The procedure we propose is based on a initial estimate $\overline{D}_D$ of $\|z^*\|_V$ (we do not require $\|z^*\|_V \leq \overline{D}_D$). It consists in successively applying the fast gradient method to a sequence of doubly smoothed dual objective functions $\theta_{\rho,\mu,\kappa}(\cdot)$ with decreasing smoothing parameter κ, restarting from scratch at each application, until condition $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \varepsilon$ is satisfied. This procedure will ensure we obtain a primal iterate $u \in S$ that is both nearly feasible for the linear constraints, i.e. $d(\mathcal{A}u, T) \leq O(\varepsilon)$, and nearly optimal for the primal objective function, i.e. $J(u) - P^* \leq O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$. Moreover, when compared to the ideal situation where the norm of the dual optimal solution is known, the total number of iterations required by this procedure only grows by a small logarithmic multiplicative factor of order $O(\max(1, \log_2\left(\frac{\|z^*\|_V}{\overline{D}_D}\right)))$.

Let $\varepsilon > 0$. The proposed global procedure goes as follows.
Initialization:
Let $\overline{D}_D$ be an initial estimate for $\|z^*\|_V$. Choose $\kappa(1) = \frac{\varepsilon}{8\overline{D}_D}$ and $j = 1$.
jth attempt
Apply (from scratch) the fast gradient method to the doubly smoothed dual function $\theta_{\rho,\mu,\kappa}(.)$ with $\mu = \frac{\varepsilon^2}{128\overline{D}_S\kappa(j)}$ and $\rho = \frac{\varepsilon^2}{128\overline{D}_T\kappa(j)}$ until the following stopping criterion is met

$$\|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} \leq \frac{\varepsilon}{4}. \tag{3.32}$$

If this last iterate satisfies

$$\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \varepsilon, \tag{3.33}$$

terminate and output $u_k = \arg\min_{u\in U}\{J(u) + \langle \mathcal{A}u, z_k\rangle + \frac{\mu}{2}\|u\|_U^2\}$.
Else let $\kappa(j+1) = \frac{1}{2}\kappa(j)$ and start the $(j+1)$th attempt.

First, we note that since the stopping criterion (3.32) checked at the jth attempt is the norm of the doubly regularized function being minimized, it is guaranteed to be met after at most

$$k = \sqrt{1 + \frac{128D_T}{\varepsilon^2} + \frac{128D_S}{\varepsilon^2}}\log\left(\frac{16\kappa(j)(1 + \frac{128D_T}{\varepsilon^2} + \frac{128D_S}{\varepsilon^2})(\theta(0) - \theta^* + \frac{\varepsilon^2}{64\kappa})}{\varepsilon}\right)$$

iterations, see bound (3.5).

We need now to estimate the number of attempts needed to reach the termination condition (3.33), and consider two cases:

1. When $\overline{D}_D$ is chosen smaller than the true norm of the dual optimal solution $\|z^*\|_V$, our global process stops for sure with criterion (3.33) when

$$\frac{\varepsilon}{16\|z^*\|_V} \leq \kappa(j) \leq \frac{\varepsilon}{8\|z^*\|_V}. \tag{3.34}$$

Indeed as soon as $\kappa(j)$ satisfies this condition we have

$$\begin{aligned}
\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} &\leq \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + \kappa\|z_k\|_V \\
&\overset{(3.6),(3.27)}{\leq} \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\sqrt{\|z^*\|_V^2 + \frac{2\mu}{\kappa}D_S + \frac{2\rho}{\kappa}D_T} \\
&\leq \|\nabla\theta_{\rho,\mu,\kappa}(z_k)\|_{V^*} + 2\kappa\|z^*\|_V \\
&\quad + 2\sqrt{2}\sqrt{\kappa\mu D_S} + 2\sqrt{2}\sqrt{\kappa\rho D_T} \\
&\overset{(3.32),(3.34)}{\leq} \varepsilon.
\end{aligned}$$

As $\kappa(1) = \frac{\varepsilon}{8\overline{D}_D}$ and $\kappa(j) = \frac{1}{2}\kappa(j-1)$, the number of attempts carried out before condition (3.34) holds satisfies $\log_2\left(\frac{\|z^*\|_V}{\overline{D}_D}\right)+1 \leq j \leq \log_2\left(\frac{\|z^*\|_V}{\overline{D}_D}\right)+2$.

When this global procedure stops, primal iterate u_k is nearly feasible

$$d(\mathcal{A}u_k, T) \leq \|\mathcal{A}u_k - x_{\rho,z_k}\|_{V^*} = \|\nabla\theta_{\rho,\mu}(z_k)\|_V^* \leq \varepsilon$$

and nearly optimal

$$\begin{aligned} J(u_k) - P^* &\overset{(3.31)}{\leq} |\langle \nabla\theta_{\rho,\mu}(z_k), z_k\rangle| + \theta^* - \theta_{\rho,\mu}(z_k) \\ &\leq \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \|z_k\|_V + \rho D_T + \mu D_S \\ &\overset{(3.6),(3.27),(3.33)}{\leq} 2\varepsilon\sqrt{\|z^*\|_V^2 + \frac{2\mu D_S}{\kappa} + \frac{2\rho D_T}{\kappa}} + \rho D_T + \mu D_S \\ &\overset{(3.34)}{\leq} \frac{25}{4}\varepsilon \|z^*\|_V . \end{aligned}$$

2. If $\overline{D}_D$ is chosen bigger than $\|z^*\|_V$, the global procedure stops after the first attempt since $\|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \frac{3\varepsilon}{4} + \frac{\varepsilon\|z^*\|_V}{4\overline{D}_D} \leq \varepsilon$, and we obtain a primal iterate u_k satisfying $d(\mathcal{A}u_k, T) \leq \|\nabla\theta_{\rho,\mu}(z_k)\|_{V^*} \leq \varepsilon$ and $J(u_k) - P^* \leq 2\varepsilon\sqrt{\|z^*\|_V^2 + 2\overline{D}_D^2} + \frac{\overline{D}_D\varepsilon}{8} \leq (2\sqrt{3} + \frac{1}{8})\overline{D}_D\varepsilon$.

In conclusion, when the global procedure stops, we have an approximately feasible solution with $d(\mathcal{A}u_k, T) \leq \varepsilon$ and a guarantee on the accuracy of the objective function of order $O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$. Although this last bound cannot usually be computed in practice, since $\|z^*\|_V$ is not known, an a posteriori guarantee on the objective accuracy can be easily computed from the right-hand side of inequality $J(u_k) - P^* \leq |\langle \nabla\theta_{\rho,\mu}(z_k), z_k\rangle| + \rho D_T + \mu D_S$.

Remark 3.5. A potential drawback of the suggested scheme is that while it allows the user to choose a guarantee ε on the feasibility of the final iterate, it does not allow the same with the objective function accuracy. Indeed, our bound for that accuracy $O(\max(\|z^*\|_V, \overline{D}_D)\varepsilon)$ involves the unknown size of the dual optimal solution[1]. However, this is unavoidable, as all known methods for convex optimization require at least some information of this type to guarantee absolute accuracy of the solution. For example, the standard analysis of the primal gradient method, which produces inequality $f(x_k) - f^* \leq \frac{Ld(x_0, X^*)}{2k}$, requires knowledge of the initial distance to the primal optimal solution set X^*, whose nature is similar to the size of the primal optimal solution.

3.8 Applications of the Double Smoothing Technique in Optimal Control

In this section, we will study optimal control problems (OCP) that can be written in the form (3.7) (more precisely in the form (3.10)). In particular, we

[1]For comparison, when D_D is known, the objective accuracy can be chosen as ϵ, and the feasibility guarantee becomes $O(\frac{\epsilon}{D_D})$, as done in the analysis of Section 3.6.

consider OCP governed by a system of linear differential equations with convex objective functional, convex constraints on the state variables at finite number of inspection moments, and point-wise convex constraints on the control variables. In order to motivate our choice of problem class, in particular the linearity of the differential equations, we first show that OCP with a nonlinear system of differential equations are NP-hard.

Consider the following OCP with convex objective function:

$$\min_{u \in L^2([0,1], R^n)} \left\{ \|x(1)\|_4^4 + \langle c, x(1) \rangle^2 : \dot{x} = -x \cdot \langle x, u \rangle + u, \; x(0) = x_0 \right\}. \tag{3.35}$$

We assume that vector c has integer coefficients.

Lemma 3.4. *Let $\|x_0\|_2^2 = 1$. Then, finding an approximate solution to problem (3.35) with absolute accuracy higher than $\hat{\epsilon} \stackrel{\text{def}}{=} \frac{1}{n(1+n^{1/2}\|c\|_1)^2}$ is NP-hard.*

Proof. In view of the system of differential equations in (3.35), we have $\langle \dot{x}, x \rangle \equiv 0$. Hence, by condition of the lemma, $\|x(t)\|_2 \equiv 1$. Note that by an appropriate control u we can move the starting point x_0 to any position at the unit sphere. Hence, the problem (3.35) is equivalent to the following finite-dimensional minimization problem:

$$\text{Find } \xi_* = \min_{\|y\|_2 = 1} \left\{ \xi(y) \stackrel{\text{def}}{=} \|y\|_4^4 + \langle c, y \rangle^2 \right\}. \tag{3.36}$$

Let us show that this problem is equivalent to solving the equation $\langle c, y \rangle = 0$ with Boolean variables (which is a well-known NP-hard problem).

If this equation has Boolean solution y_* with coefficients $y_*^{(i)} = \pm m^{-1/2}$, $i = 1, \ldots, n$, then $\xi_* = \frac{1}{m}$. On the other hand, note that for any $y \in R^m$ with unit Euclidean norm we have $\|y\|_4^4 = \frac{1}{m} + \sum\limits_{i=1}^{m} \left((y^{(i)})^2 - \frac{1}{m} \right)^2$. If we manage to find such a point y with $\xi(y) - \xi_* < \hat{\epsilon}$, then in the case $\xi(y) \geq \frac{1}{m} + \hat{\epsilon}$ we guarantee the absence of Boolean solutions. If $\xi(y) < \frac{1}{m} + \hat{\epsilon}$, then $|\langle c, y \rangle| < \hat{\epsilon}^{1/2}$ and $\max\limits_{1 \leq i \leq m} \left| (y^{(i)})^2 - \frac{1}{m} \right| < \hat{\epsilon}^{1/2}$. In this case, we can define the Boolean vector $u^{(i)} = \frac{1}{m^{1/2}} \cdot \text{sign}(y^{(i)})$, $i = 1, \ldots, m$. For this vector we have

$$\begin{aligned} |\langle c, u \rangle| &= |\langle c, y \rangle + \langle c, u - y \rangle| \\ &\leq |\langle c, y \rangle| + \left| \sum_{i=1}^{m} c^{(i)} \cdot \text{sign}(y^{(i)}) \left(\tfrac{1}{n^{1/2}} - |y^{(i)}| \right) \right| \\ &< \hat{\epsilon}^{1/2} + \|c\|_1 \max_{1 \leq i \leq m} \left| \tfrac{1}{m^{1/2}} - |y^{(i)}| \right| \\ &< \hat{\epsilon}^{1/2} (1 + m^{1/2} \|c\|_1) = \tfrac{1}{m^{1/2}}. \end{aligned}$$

Since vector c has integer coefficients, we conclude that $\langle c, u\rangle = 0$. $\square$

3.8.1 Class of optimal control problems and reformulation

Consider the following optimal control problem:

$$\inf_u \{\int_0^1 F(t,u(t))dt : \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \\ x(t_i) \in T_i \quad i = 1..., N, \\ u(t) \in S(t) \quad \text{a.e in } [0,1]\}, \tag{3.37}$$

with variables $x(t) \in R^n$ and $u(t) \in R^m$, $t \in [0,1]$ and where sets $S(t) \subset R^m$, $t \in [0,1]$, are closed and convex with bounded graph $\overline{S} \stackrel{\text{def}}{=} \cup_{t\in[0,1]} S(t)$. We assume that function $F : [0,1] \times \overline{S} \to R$ is bounded and continuously differentiable and convex in the second argument. We measure control variables with the norm $\|u\|_2^2 = \int_0^1 \|u(t)\|_2^2 dt$. We assume that $A(t) \in \mathcal{C}([0,1], R^{n\times n})$ and $B(t) \in \mathcal{C}([0,1], R^{n\times m})$. In problem (3.37), we have a finite number of inspection moments $t_i \in (0,1]$, $i = 1,\dots,N$, and we assume that each subset $T_i \subset R^n$ is a closed bounded convex set. Let us rewrite the problem (3.37) in terms of control u. Denote by $\Phi(t,\tau)$ the transition matrix of the system, i.e. the unique solution of the following matrix Cauchy problem

$$\tfrac{d}{dt}\Phi(t,\tau) \quad = \quad A(t)\Phi(t,\tau), \quad t \geq \tau, \quad \Phi(\tau,\tau) = I.$$

Remark 3.6. When the system is time-invariant, i.e. $A(t) = A$ and $B(t) = B$ for all $t \in [0,1]$, transition matrix Φ is the usual matrix exponent

$$\Phi(t,\tau) \quad = \quad e^{(t-\tau)A} = I + \sum_{k=1}^{\infty} \tfrac{A^k (t-\tau)^k}{k!}.$$

From classical optimal control theory (e.g [40]), we know that the state trajectory $x(t)$, generated by the system of differential equations under the control $u(t)$, is defined by $x(t) = \Phi(t,0)x_0 + \int_0^t \Phi(t,\tau)B(\tau)u(\tau)d\tau$, $t \in [0,1]$. Therefore, the constraint $x(t_i) \in T_i$ can be expressed as $\mathcal{A}_i(u) \in \bar{T}_i$ for each $i = 1,\dots,N$ using

$$\mathcal{A}_i(u) \quad \stackrel{\text{def}}{=} \quad \int_0^{t_i} \Phi(t_i,\tau)B(\tau)u(\tau)d\tau \quad \text{and} \quad \overline{T}_i \stackrel{\text{def}}{=} T_i - \Phi(t_i,0)x_0, \tag{3.38}$$

where $\Phi(t_i,0)x_0$ is the value at time t_i of the unique solution of Cauchy problem

$$\dot{x}(t) = A(t)x(t), \quad x(0) = x_0.$$

Linear operator $\mathcal{A}_i : L^2([0,1], R^m) \to R^n$ can also be written as

$$\mathcal{A}_i u = \int_0^1 A_i(\tau)u(\tau)d\tau \text{ where } A_i(\tau) \stackrel{\text{def}}{=} \begin{cases} \Phi(t_i,\tau)B(\tau), & \text{when } \tau \in [0,t_i], \\ 0, & \text{when } \tau \in]t_i,1]. \end{cases}$$

Defining now $J(u) = \int_0^1 F(t,u(t))dt$ and $S = \{u \in L^2([0,1],R^m) : u(t) \in S(t) \text{ a.e. in } [0,1]\}$, the optimal control problem (3.37) can be rewritten in the form (3.8), and therefore in the form (3.7), if we define the convex set $T = \bar{T_1} \times \bar{T_2} \times ... \times \bar{T_N} \subset R^{N\times n}$ and the linear operator $\mathcal{A} : L^2([0,1],R^m) \to R^{N\times n}$ such that $\mathcal{A}u : R^{N\times n} \to R$ satisfies $\langle \mathcal{A}u, z\rangle = \sum_{i=1}^N \langle \mathcal{A}_i u, z_i\rangle$ for all $u \in L^2([0,1],R^m)$ and $z = (z_1,...,z_N) \in R^n$. Hence, we can solve it by the double smoothing technique. This approach assumes that we are able to solve the following problems easily in a pointwise way

$$\max_{u\in S(t)} \left\{ -F(t,u) - \sum_{i=1}^N \langle u, A_i(t)^T z^i\rangle - \frac{\mu}{2}\|u\|_2^2 \right\},$$

where $A_i(t)$ depends directly on the state transition matrix. However, in practice the state transition matrix $\Phi(t_i,t)$ is often not known explicitly. Instead, we can compute the function $A_i(t)^T z^i$ as a solution of some system of differential equations. Indeed, we have (see e.g. Theorem 1.2 in [42]) $\frac{d}{dt}\Phi^T(t_i,t) = -A(t)^T\Phi^T(t_i,t)$. Therefore $\Phi(t_i,t)^T$ is the state transition matrix of the system $\dot{v}(t) = -A(t)^T v(t)$. Hence, $A_i(t)^T z^i = B(t)^T v(t)$, where $v(t)$ is the unique solution of Cauchy problem

$$\dot{v}(t) \quad = \quad -A(t)^T v(t), \quad v(t_i) = z^i, \quad t \in [0,t_i], \tag{3.39}$$

extended by zero for $t \in [t_i, 1]$.

Remark 3.7. At first glance, it seems that we are restricted to objective functionals depending only on the control $u(t)$ and not on the state variable $x(t)$. In fact, using the state transition matrix, we can also consider any convex functions depending on linear functionals of the state. Such a functional can be defined as

$$\begin{aligned} l(x) &= \int_0^1 \langle x(t), a(t)\rangle dt = \int_0^1 \langle \int_0^t \Phi(t,\tau)B(\tau)u(\tau)d\tau, a(t)\rangle dt \\ &= \int_0^1 \int_0^t \langle \Phi(t,\tau)B(\tau)u(\tau), a(t)\rangle d\tau dt \\ &= \int_0^1 \int_0^t \langle u(\tau), B(\tau)^T\Phi(t,\tau)^T a(t)\rangle d\tau dt \\ &= \int_0^1 \int_\tau^1 \langle u(\tau), B(\tau)^T\Phi(t,\tau)^T a(t)\rangle dt d\tau \stackrel{\text{def}}{=} \int_0^1 \langle u(\tau), h(\tau)\rangle d\tau, \end{aligned}$$

with $h(\tau) = \int_\tau^1 B(\tau)^T \Phi(t,\tau)^T a(t) dt$. Another possibility is as follows:

$$\begin{aligned} l(x) &= \langle x(t_i), a \rangle = \langle \int_0^{t_i} \Phi(t_i,\tau) B(\tau) u(\tau) d\tau, a \rangle \\ &= \int_0^{t_i} \langle \Phi(t_i,\tau) B(\tau) u(\tau), a \rangle d\tau \overset{\text{def}}{=} \int_0^{t_i} \langle u(\tau), h(\tau) \rangle d\tau, \end{aligned}$$

with $h(\tau) = B(\tau)^T \Phi(t_i,\tau)^T a$.

3.8.2 Evaluation of $\|\mathcal{A}_i\|_2$

In order to solve the primal-dual problem (3.7)-(3.9) by a double smoothing technique, we need to bound the norm $\|\mathcal{A}\|_2 \leq [\sum_{i=1}^N \|\mathcal{A}_i\|_2^2]^{1/2}$. Moreover, from the estimates (3.25), (3.26), it is clear that this norm is an essential element in the global complexity bound of our problem. In this section, we first derive a closed-form representation for the norm $\|\mathcal{A}_i\|_2$ using the reachability Gramian of the dynamical system. However, this quantity is not easily computable (it needs the knowledge of the transition matrix). Moreover, its dependence in the length of time interval is not very transparent. Therefore, in the next section, we obtain some simple upper bounds for the norms $\|\mathcal{A}_i\|_2$, which can be easily computed by solving Linear Matrix Inequalities (LMI).

Let us derive first the exact expression for $\|\mathcal{A}_i\|_2$. By definition,

$$\|\mathcal{A}_i\|_2 = \sup_{u \in L^2([0,1],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \ \|u\|_{L^2([0,1],R^m)} = 1 \right\}.$$

Since the vector $\mathcal{A}_i u$ does not depend on values of $u(t)$ for $t \in (t_i, 1]$, we can consider the restriction of $\mathcal{A}_i$ on $L^2([0,t_i],R^m) : u \to \int_0^{t_i} \Phi(t_i,\tau) B(\tau) u(\tau) d\tau$. Then

$$\|\mathcal{A}_i\|_2 = \sup_{u \in L^2([0,t_i],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \ \|u\|_{L^2([0,t_i],R^m)} = 1 \right\},$$

and the operator $\mathcal{A}_i^*$ maps $y \in R^n$ into function $B(t)^T \Phi(t_i,t)^T y \in L^2([0,t_i]$.

For all $t_i > 0$, $i = 1, \ldots, N$, define the reachability Gramians

$$W_r(0,t_i) = \int_0^{t_i} \Phi(t_i,\tau) B(\tau) B(\tau)^T \Phi(t_i,\tau)^T d\tau = \mathcal{A}_i \mathcal{A}_i^*,$$

which are symmetric positive semidefinite matrices. Recall the following definition:

Definition 3.1. The system

$$\dot{x}(t) \quad = \quad A(t)x(t) + B(t)u(t), \quad x(0) = 0 \tag{3.40}$$

is called reachable on $[0, \hat{t}]$ if for any $\hat{x} \in R^n$ there exist a control $u(t)$ such that $x(\hat{t}) = \hat{x}$.

Reachability is closely related to the reachability Gramian (e.g. Corollary 2.3 in [1]).

Theorem 3.5. *The system (3.40) is reachable on $[0, \hat{t}]$ if and only if the Gramian $W_r(0, \hat{t})$ is positive definite.*

Let us come back now to the definition of the norm $\|\mathcal{A}_i\|_2$. We have:

$$\begin{aligned} \|\mathcal{A}_i\|_2 &= \sup_{u \in L^2([0,t_i],R^m)} \left\{ \|\mathcal{A}_i u\|_2 : \ \|u\|_{L^2([0,t_i],R^m)} = 1 \right\} \\ &= \left[\inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} \right]^{-1}. \end{aligned}$$

If the system is reachable on $[0, t_i]$, then Im $\mathcal{A}_i(L^2([0, t_i], R^m)) = R^n$, and we have

$$\begin{aligned} & \inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} \\ = & \inf_{\substack{x_i \in R^n, \|x_i\|_2 = 1 \\ u \in L^2([0,t_i],R^m)}} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \mathcal{A}_i u = x_i \right\}. \end{aligned}$$

In order to solve this minimization problem, we will use the following simple result:

Lemma 3.6. *Let H be a Hilbert space and the linear operator $A : H \to R^L$ be nondegenerate: $AA^* \succ 0$. Then for any $b \in R^L$ and $f \in H$, the Euclidean projection $\pi_b(f)$ of f onto the subspace $\mathcal{L}_b = \{g \in H : Ag = b\}$ corresponds to $\pi_b(f) = f + A^*(AA^*)^{-1}(b - Af)$.*

Therefore

$$\inf_{\substack{u \in L^2([0,t_i],R^m), \\ \mathcal{A}_i u = x_i}} \|u\|_2 = \|\mathcal{A}_i^*(\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i\|_2 \ = \langle (\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i, x_i\rangle^{1/2}.$$

Therefore,

$$\begin{aligned} \inf_{u \in L^2([0,t_i],R^m)} \left\{ \|u\|_{L^2([0,t_i],R^m)} : \ \|\mathcal{A}_i u\|_2 = 1 \right\} &= \inf_{\|x_i\|_2 = 1} \langle (\mathcal{A}_i\mathcal{A}_i^*)^{-1}x_i, x_i\rangle^{1/2} \\ &= \lambda_{\min}^{1/2}((\mathcal{A}_i\mathcal{A}_i^*)^{-1}), \end{aligned}$$

and we conclude that

$$\|\mathcal{A}_i\|_2 \quad = \quad \lambda_{\min}^{-1/2}((\mathcal{A}_i\mathcal{A}_i^*)^{-1}) \ = \ \lambda_{\max}^{1/2}(\mathcal{A}_i\mathcal{A}_i^*),$$

where $\mathcal{A}_i\mathcal{A}_i^* = W_r(0, t_i)$ is the reachability Gramian.

3.8.3 Bounding the growth of norms $\|\mathcal{A}_i\|_2$ with time

In the previous section, we have shown that the norm $\|\mathcal{A}_i\|_2$ is equal to the square root of the maximal eigenvalue of the reachability Gramian on the interval $[0, t_i]$. Simple examples show that this norm can grow exponentially with t_i. However, for stable systems, the situation is much better.

In this section, we derive the bounds for the growth of the norms $\|\mathcal{A}_i\|_2$ from the stability characteristics of the linear time-varying system:

$$\dot{x}(t) \quad = \quad A(t)x(t), \quad t \geq 0, \tag{3.41}$$

where the matrix $A(t)$ depends continuously on time.

Recall that the state $x = 0$ is always an equilibrium of the system (3.41). It is the unique equilibrium if $A(t)$ is nonsingular for all $t \geq 0$. The following facts are standard (e.g. [1]):

Theorem 3.7. *The equilibrium $x = 0$ is stable if and only if the solutions of the linear systems are bounded, that is $\sup_{t\geq\tau} \|\Phi(t,\tau)\|_2 \stackrel{\text{def}}{=} k(\tau) < \infty, \quad \forall \tau \geq 0$. It is uniformly stable if and only if $\sup_{\tau\geq 0} k(\tau) = \sup_{\tau\geq 0} \sup_{t\geq\tau} \|\Phi(t,\tau)\|_2 \stackrel{\text{def}}{=} \kappa_0 < \infty$. Finally, it is is exponentially stable if $\int_0^t \|\Phi(t,\tau)\|_2^2 \, d\tau \leq C$ for all $t \geq 0$ where constant C is independent on t.*

Using these stability results, we can obtain some estimates for the growth of $\|\mathcal{A}_i\|_2$.

Theorem 3.8. *If the equilibrium $x = 0$ is stable and $k_1 \stackrel{\text{def}}{=} \sup_{t\geq 0} \|B(t)\|_2 < \infty$, then*

$$\|\mathcal{A}_i\|_2 \quad \leq \quad k_1 \left[\int_0^{t_i} k^2(\tau) d\tau \right]^{1/2}. \tag{3.42}$$

Proof. For all $u \in L^2([0, t_i]R^m)$, we have

$$\begin{aligned} \|\mathcal{A}_i u\| \quad &= \quad \left\| \int_0^{t_i} \Phi(t_i,\tau) B(\tau) u(\tau) d\tau \right\| \quad \leq \quad \int_0^{t_i} \|\Phi(t_i,\tau) B(\tau)\|_2 \, \|u(\tau)\|_2 \, d\tau \\ &\leq \quad \left[\int_0^{t_i} \|\Phi(t_i,\tau)\|_2^2 \|B(\tau)\|_2^2 d\tau \cdot \int_0^{t_i} \|u(\tau)\|_2^2 d\tau \right]^{1/2} \\ &\leq \quad k_1 \left[\int_0^{t_i} k^2(\tau) d\tau \right]^{1/2} \|u\|_2 , \end{aligned}$$

hence inequality (3.42). □

This upper bound depends on the growth of integral $\int_0^{t_i} k^2(\tau)d\tau$ with respect to t_i, which can be very fast. Moreover, it can happen that function $k(\cdot)$ does not belong to $L^2([0,t_i])$, in which case bound (3.42) gives no information. However, if we assume uniform stability of the equilibrium $x=0$, we can get much better bounds.

Theorem 3.9. *If equilibrium $x=0$ is uniformly stable and $k_1<\infty$, then $\|\mathcal{A}_i\|_2 \le k_0 k_1 \sqrt{t_i}$.*

The proof of this theorem is the same as that of Theorem 3.8. However, now we can ensure a sublinear bound for the growth $\|\mathcal{A}_i\|_2$ with respect to t_i. If we strengthen again the stability assumption, we can obtain an upper bound independent on t_i.

Theorem 3.10. *Let equilibrium $x=0$ be exponentially stable and $k_1<\infty$. Then $\|\mathcal{A}_i\|_2 \le k_1\sqrt{C}$.*

Again, this fact can be easily derived from the arguments of the proof of Theorem 3.8. In some cases, we can obtain a computable upper bound for the norm $\|\mathcal{A}_i\|_2$. Recall the following well-known sufficient condition for the global exponential stability.

Theorem 3.11. *[1] Let the linear system (3.41) be time-invariant, and assume there exists a matrix $P=P^T\succ 0$ such that $A^TP+PA\prec 0$. Then equilibrium $x=0$ is globally exponentially stable.*

Under conditions of this theorem, there exists $\eta_1>0$ such that the following LMI

$$A^TP+PA\preceq -\eta_1 P,\quad P=P^T\succ 0,$$

admits a solution. Matrix P and constant η_1 can help us to obtain an explicit upper bound for the norm $\|\mathcal{A}_i\|_2$. Indeed, by definition, $\mathcal{A}_i u$ is the position at time t_i of the point $x(t)$ of the unique trajectory defined by the linear system

$$\dot{x}(t) \;=\; Ax(t)+Bu(t),\quad x(0)=0.$$

Therefore, we have $\|x(t_i)\|_2^2=\langle x(t_i),x(t_i)\rangle \le \frac{\langle Px(t_i),x(t_i)\rangle}{\lambda_{\min}(P)}=\frac{R(t_i)}{\lambda_{\min}(P)}$, where we have introduced function $R(t)\stackrel{\rm def}{=}\langle Px(t),x(t)\rangle$. Its derivative can be bounded as follows:

$$\begin{array}{rcl}
\dot{R}(t) &=& \langle P,x(t)\dot{x}(t)^T+\dot{x}(t)x(t)^T\rangle \\
&=& \langle P,x(t)(Ax(t)+Bu(t))^T+(Ax(t)+Bu(t))x(t)^T\rangle \\
&=& \langle P(Ax(t)+Bu(t)),x(t)\rangle+\langle Px(t),Ax(t)+Bu(t)\rangle \\
&=& \langle (PA+A^TP)x(t),x(t)\rangle+2\langle Px(t),Bu(t)\rangle \\
&\le& -\eta_1\langle Px(t),x(t)\rangle+2\langle Px(t),Bu(t)\rangle \;\le\; \frac{1}{\eta_1}\langle PBu(t),Bu(t)\rangle.
\end{array}$$

Since $x(0) = 0$, we get

$$\begin{aligned} R(t_i) &= \int_0^{t_i} \dot{R}(t)dt \;\leq\; \tfrac{1}{\eta_1} \int_0^{t_i} \langle PBu(t), Bu(t)\rangle dt \\ &\leq \tfrac{1}{\eta_1}\lambda_{\max}(P) \int_0^{t_i} \|Bu(t)\|_2^2\, dt \;\leq\; \tfrac{1}{\eta_1}\lambda_{\max}(P)\, \|B\|_2^2\, \|u\|_2^2 . \end{aligned}$$

Hence, $\|\mathcal{A}_i u\|_2^2 \leq \frac{\lambda_{\max}(P)}{\eta_1 \lambda_{\min}(P)}\, \|B\|_2^2\, \|u\|_2^2$, and therefore $\|\mathcal{A}_i\|_2^2 \leq \frac{\lambda_{\max}(P)}{\eta_1 \lambda_{\min}(P)}\, \|B\|_2^2$. If we want to obtain the best upper bound for $\|\mathcal{A}_i\|_2$, we need to solve the following optimization problem in the variables η_1, η_2, η_3, and P:

$$\min \left\{ \tfrac{\eta_3}{\eta_1 \eta_2} : \quad A^T P + PA \preceq -\eta_1 P, \quad \eta_2 I \preceq P \preceq \eta_3 I, \quad \eta_1, \eta_2, \eta_3 \geq 0 \right\}. \tag{3.43}$$

Although this problem is non-convex, we can provide an upper bound for its optimal value that is computable with a convex LMI. Indeed, we note that

$$\|\mathcal{A}_i\|_2^2 \leq \min \left\{ \frac{\eta_3}{\eta_1 \eta_2} : \quad A^T P + PA \preceq -\eta_1 \eta_3 I, \quad \eta_2 I \preceq P \preceq \eta_3 I, \quad \eta_1, \eta_2, \eta_3 \geq 0 \right\}$$

since the feasible set in the right-hand side is smaller than that of (3.43). Furthermore, if we introduce new variables: $\tilde{P} = \frac{P}{\eta_1 \eta_3}$, $\tilde{\eta}_2 = \frac{\eta_2}{\eta_1 \eta_3}$ and $\tilde{\eta}_3 = \frac{1}{\eta_1}$, we obtain a convex problem that can be solved in polynomial time

$$\min \left\{ \frac{\tilde{\eta}_3^2}{\tilde{\eta}_2} : A^T \tilde{P} + \tilde{P} A \preceq -I, \quad \tilde{\eta}_2 I \leq P \leq \tilde{\eta}_3 I, \quad \tilde{\eta}_2, \tilde{\eta}_3 \geq 0 \right\}.$$

3.9 Comparison with the literature and conclusion

The subject of this chapter can be summarized as the development of an efficient first-order method (obtained using the double smoothing technique) to solve partially finite (or finite) convex optimization problems with linear constraints.

Partially finite convex problems have been extensively studied in a theoretical way with duality results, weak constraint qualification ([13, 14, 16, 17, 31, 32]) and applications for example to maximum entropy ([12, 15]).

On the other hand, it is not the first time that the smoothing technique is used for solving finite-dimensional convex problems with linear constraints by first-order methods. As our approach can be also interesting for solving finite-dimensional problems, we briefly mention these results here, and discuss the

differences with our approach.

In [43], the authors consider the case of a conic problem with linear objective function, i.e. $J(u) = \langle c^*, u\rangle$, with $c^* \in U^*$, $T = \{b^*\} \subset V^*$ and $S = \mathcal{L} \subset U$ a closed convex cone. Using the rich duality theory for such kind of conic problems, they consider a primal-dual approach. The main idea is to reformulate the primal dual optimality conditions $\mathcal{A}^* y + s^* - c^* = 0, \mathcal{A}u - b^* = 0, \langle c^*, u\rangle - \langle b^*, y\rangle = 0, (u, y, s^*) \in \mathcal{L} \times V \times \mathcal{L}^*$ as a nonsmooth convex problem:

$$\overline{f} = \min_{z\in Z}\{f(z) = \|Ez - e\|_*\} = \min_{z\in Z} \max_{\|w\|\leq 1} \langle Ez - e, w\rangle \tag{3.44}$$

where $\|.\|$ denotes a norm on $U \times V \times R$, $\|\cdot\|_*$ its dual norm, $z = (u, y, s^*)^T$, $e = (c^*, b^*, 0)^T$,

$$E = \begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ c^* & -b^* & 0 \end{pmatrix}$$

and $Z = \mathcal{L} \times V \times \mathcal{L}^*$. Applying the smoothing technique to the function f, they are able to find a primal-dual solution z_ϵ such that $\|Ez_\epsilon - e\|_* \leq \epsilon$ in $O\left(\frac{1}{\epsilon}\right)$. This primal-dual approach does not work in our case for two reasons. First, primal-dual optimality conditions cannot be expressed as a linear system $Ez = e$ (subject to a conic constraint). Furthermore, we do not want to work in the primal-dual space but preferably in the dual one due to our asymmetry assumption (the problem (3.44) can be infinite-dimensional in our framework).

The approach considered in [6] is more comparable to what we are doing. Their problem class is composed of problems of the form (3.7) with J a (not necessarily smooth) convex function, $S = U$ and $T = \mathcal{K} - b$, where $\mathcal{K}$ is a closed convex cone. Dualizing the constraint $\mathcal{A}u + b \in \mathcal{K}$, they obtain a dual problem with conic constraint: $\max_{z\in\mathcal{K}^*} g(z)$ where $g(z) = \inf_u J(u) - \langle \mathcal{A}u + b, z\rangle$. They apply the smoothing technique to the dual objective function and compare different optimal first-order methods of smooth convex optimization for solving the smoothed dual problem. They are able to solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\right)$ iteration. To reconstruct a nearly optimal primal solution u_ϵ from a nearly optimal dual solution z_ϵ, they suggest to choose u_ϵ as the minimizer of the optimization subproblem defining the smoothed dual objective function at the point z_ϵ. However, this suggestion is not supported by the analysis of the convergence rates (see our discussion at the end of Section 3.4.1).

In the framework of separable convex problems, the smoothing technique has also been applied in [49] to convex problems with linear coupling constraint. Dualizing the coupling constraint, the authors obtain a dual objective function that can be computed in a separable way. Applying a simple smoothing to this

dual objective function, they obtain a smooth dual objective function keeping the separability structure. Here also, this approach allows them to solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\right)$ iterations. Averaging of the minimizers of the subproblems defining the smoothed dual objective function at the different dual iterates is proposed to reconstruct a primal solution. It is proved that the quality of this primal solution is also of order ϵ. It also depends on the norm of the dual optimal solution (which is typically unknown).

The approach considered in this chapter also allows us to exploit the separability structure of decomposable problems with linear coupling constraints (see the two examples given in Section 3.3). In our work, we apply a double smoothing that gives us a possibility to reconstruct more easily a nearly optimal primal solution from a nearly optimal dual solution, without averaging. The price that we pay for this simplicity is a logarithmic term $\log\left(\frac{1}{\epsilon}\right)$ in the complexity. For the level of accuracy we are interested in, the logarithmic factor is not distinguishable from an absolute constant. Furthermore, whereas we use also in our analysis the norm of the dual optimal solution, we are able to provide an explicit upper bound for this quantity.

More generally, to the best of our knowledge, this work is the first one where the smoothing technique is applied for solving infinite-dimensional problems. In particular, the double smoothing technique can be applied to optimal control problems governed by a system of linear differential equations with the constraint that the trajectory crosses some convex set at certain moments of time (see section 3.8).

Part II

First-order methods with inexact information

Chapter 4

First-Order Methods with Inexact Oracle: the smooth convex case

This chapter corresponds to the paper [24]:
O. Devolder, F. Glineur and Yu. Nesterov. **First-order Methods of Smooth Convex Optimization with Inexact Oracle.** *Mathematical Programming, Serie A, Accepted*, (2013).

Chapter 4 in four Questions/Answers.

◇ *What is the effect of (deterministic) inexact first-order information on existing first-order methods of smooth convex optimization, namely the Primal Gradient Method (PGM), the Dual Gradient Method (DGM) and the Fast Gradient Method (FGM) ?*

Two very different behaviors appear. On one side, the PGM/DGM are slow but robust. They have a slow convergence rate proportional with $O(\frac{1}{k})$ but do not accumulate the errors done at each iteration. On the other side, the FGM is fast but sensitive to errors. It exhibits a fast convergence rate proportional to $O(\frac{1}{k^2})$ but suffers from an accumulation of errors at a linear rate $k\delta$, where δ is the individual error in each first-order information.

◇ *Is it possible to develop a first-order method which is at the same time fast and robust with respect to oracle errors ? Or is there an intrinsic link between the fastness of a method and its sensitivity with respect to errors ?*

There is no free lunch and no hope to develop a perfect method which would be as fast as the FGM and as robust as the PGM/DGM. The fastness of a first-order method and its sensitivity to errors are linked: the faster a method is, the higher its sensitivity to errors must be also. Accumulation of errors of the FGM does not come from our analysis or from a specific problem of this method, it is an intrinsic property of any fast first-order method.

◇ *Is it possible to apply first-order methods, initially designed for smooth convex problems, to objective functions with a weaker level of smoothness ?*

Yes! The notion of (δ, L)-oracle, that we introduce in order to model errors in the first-order information, can be also used in order to represent a lack of smoothness. The exact oracle of a nonsmooth or a weakly smooth convex function can be seen as a particular case of (δ, L)-oracle. As a consequence, the first-order methods initially developed for smooth convex problems (i.e. the PGM/DGM and FGM) can be also applied to nonsmooth and weakly smooth convex functions, obtaining in some sense universal first-order methods. These results break the wall that seems at first sight to exist between smooth and nonsmooth convex optimization.

◇ *What is the behavior of the PGM/DGM and FGM when applied to a nonsmooth or weakly smooth function ? How to choose the stepsizes, taking into account the lack of smoothness, in a optimal way ? In particular, is it possible to obtain a universal optimal first-order method, exhibiting an optimal complexity on smooth, weakly smooth and nonsmooth convex problems ?*

The PGM/DGM are only optimal in the nonsmooth case where the PGM is nothing else that the subgradient/mirror-descent method. On the other hand, the FGM, when used with well-chosen stepsizes (depending on the level of smoothness of the objective function), can be seen as a universal optimal first-order method, reaching the optimal complexity in the smooth, weakly smooth and nonsmooth cases.

For the detailed choice of the stepsizes and the corresponding complexities, we refer the reader to section 4.7.

Contents

4.1	The (δ, L)-oracle	**104**
4.1.1	Motivation and definition	104
4.1.2	Properties	105
4.1.3	Examples	111
4.2	Functions defined by an optimization subproblem	**116**
4.2.1	Accuracy measures for approximate solutions	116
4.2.2	Functions obtained by smoothing techniques	118
4.2.3	Moreau-Yosida regularization	119
4.2.4	Functions defined by Augmented Lagrangians	120
4.3	Gradient Methods with (δ, L)-oracle	**122**
4.3.1	Primal gradient method	122
4.3.2	Dual gradient method	123
4.4	Fast Gradient Method with (δ, L)-oracle	**125**
4.4.1	Convergence analysis	125
4.4.2	Error accumulation	128
4.5	Comparison Classical and Fast Gradient Methods	**129**
4.5.1	Oracle accuracy δ can be freely chosen	130
4.5.2	Oracle accuracy δ is fixed.	130
4.6	Comparison with other types of inexact oracle	**133**
4.7	Application to functions with lack of smoothness	**135**
4.7.1	Solving weakly smooth problems	135
4.7.2	Solving composite optimization problems	138
4.8	First-order methods and error accumulation	**139**

Standard analysis of first-order methods assumes availability of exact first-order information. Namely, the oracle must provide at each given point the exact values of the function and its gradient. However, in many convex problems, including those obtained by smoothing techniques, the objective function and its gradient are computed by solving another auxiliary optimization problem. In practice, we are often only able to solve these subproblems approximately. Hence, in that context, numerical methods solving the outer problem are provided with inexact first-order information. This led us to investigate the behavior of first-order methods working with an inexact oracle.

We introduce in Section 4.1 a new definition of inexact first-order oracle and list a few simple examples. In Section 4.2, we show how our concept is applicable to situations when the inexact oracle is computed by an auxiliary optimization problem. In particular, we consider convex-concave saddle point problems, augmented Lagrangians, and Moreau-Yosida regularization.

In Sections 4.3 and 4.4, we consider classical (primal and dual) and fast gradient methods, designed for the class of convex functions with Lipschitz-continuous gradient. We obtain efficiency estimates when these methods are used with an inexact first-order oracle. We also study the link between desired accuracy for the objective function and necessary accuracy for the oracle. We observe that the superiority of the fast gradient methods over the classical ones is no longer absolute when an inexact oracle is used, because FGMs suffer from error accumulation. In particular, fast methods require first-order information with higher accuracy than standard gradient methods to obtain a solution with a given accuracy. Therefore, the choice between these methods depends on the availability and relative cost of an inexact oracle at different levels of accuracy, as is explained in Section 4.5.

In Section 4.6, we compare our approach with other definitions of inexact oracle, as applied to the smoothed max-representable functions typically obtained by the smoothing techniques [21, 2]. We show that our definition can give better complexity results.

Our definition of inexact oracle is applicable to nonsmooth and weakly smooth convex problems. Section 4.7 shows how to apply first-order methods designed for smooth convex optimization to functions with a weaker level of smoothness. For that, we show that (exact) first-order information for a nonsmooth problem, such as subgradients, can be viewed as an inexact oracle, so that the methods of Sections 4.3 and 4.4 can be applied. We obtain in this way "universal" first-order methods possessing optimal rates of convergence for objective functions with different levels of smoothness.

Finally, in Section 4.8, we obtain lower bounds on the rate of error accumulation for any first-order method using an inexact oracle, which shows that all methods discussed in this chapter have the lowest possible rate of error accumulation. In particular, it appears that while slower standard gradient methods are able to maintain an error comparable to the oracle accuracy, any optimal method must suffer from error accumulation.

4.1 The (δ, L)-oracle

4.1.1 Motivation and definition

Consider $F_L^{1,1}(Q)$, the class of convex functions on convex set Q whose gradient is Lipschitz-continuous with constant L. It is well-known that functions

belonging to this class satisfy

$$0 \;\leq\; f(x) - \big(f(y) + \langle \nabla f(y), x - y\rangle\big) \;\leq\; \frac{L}{2}\,\|x-y\|_E^2 \;\text{ for all } x, y \in Q, \quad (4.1)$$

see the top of Figure 1. Moreover, it is easy to check that, for a given y, quantities $f(y)$ and $\nabla f(y)$ are uniquely determined by this pair of inequalities. Therefore, membership in $F_L^{1,1}(Q)$ can be characterized by the existence of an oracle returning for each point $y \in Q$ a pair $(f_L(y), g_L(y)) \in \mathbb{R} \times E^*$, necessarily equal to $(f(y), \nabla f(y))$, satisfying

$$0 \;\leq\; f(x) - \big(f_L(y) + \langle g_L(y), x - y\rangle\big) \;\leq\; \frac{L}{2}\,\|x-y\|_E^2 \;\text{ for all } x \in Q$$

(both zeroth-order and first-order information are included in the oracle). Our definition of an inexact oracle simply consists in introducing a given amount δ of tolerance in this pair of inequalities (see bottom of Figure 1).

Definition 4.1. Let function f be convex on convex set Q. We say that it is equipped with a first-order (δ, L)-oracle if for any $y \in Q$ we can compute a pair $(f_{\delta,L}(y), g_{\delta,L}(y)) \in \mathbb{R} \times E^*$ such that

$$0 \;\leq\; f(x) - \big(f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y\rangle\big) \leq\; \frac{L}{2}\,\|x-y\|_E^2 + \delta \;\text{ for all } x \in Q. \quad (4.2)$$

Constant δ will be called the accuracy of the oracle. A function f belongs to $F_L^{1,1}(Q)$ if and only if it admits a $(0, L)$-oracle, namely $(f_{0,L}(y), g_{0,L}(y)) = (f(y), \nabla f(y))$. However, the class of functions admitting a (δ, L)-oracle is strictly larger, and includes nonsmooth functions, as we will see later.

4.1.2 Properties

We list here a few important properties of (δ, L)-oracles.

- ◇ A (δ, L)-oracle provides a lower δ-approximation of the function value. Indeed, taking $x = y$ in (4.2), we obtain

$$f_{\delta,L}(y) \leq f(y) \leq f_{\delta,L}(y) + \delta. \quad (4.3)$$

- ◇ A (δ, L)-oracle provides a δ-subgradient of f at $y \in Q$, i.e.

$$g_{\delta,L}(y) \in \partial_\delta f(y) = \{z \in E^* : f(x) \geq f(y) + \langle z, x - y\rangle - \delta \quad \forall x \in Q\}.$$

  Indeed, using the first inequality in (4.2) and (4.3), we have for all $x, y \in Q$

$$f(x) \;\geq\; f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y\rangle \;\geq\; f(y) + \langle g_{\delta,L}(y), x - y\rangle - \delta. \quad (4.4)$$

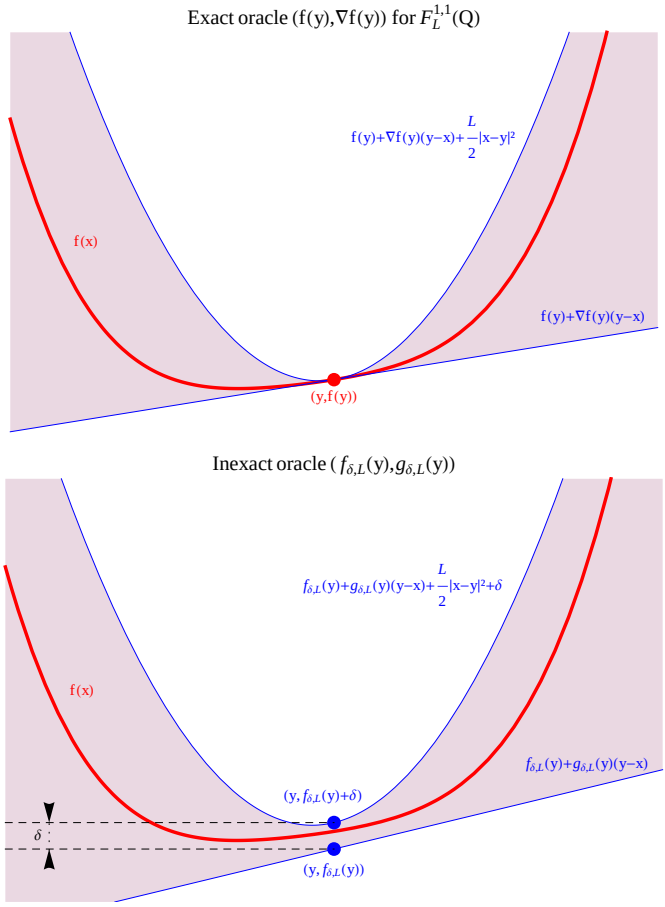


Figure 1: Illustration of lower and upper bounds (blue lines) implied by the definition of an exact (top) and inexact (bottom) oracle.

Methods of nonsmooth convex optimization based on δ-subgradients have a long history (see e.g. [74, 68, 20, 50] for subgradient methods, and [20, 29, 41] for proximal point and bundle methods). We will show later that a standard subgradient can also satisfy the second inequality in (4.2), which opens the possibility of using the concept of inexact oracle in the context of nonsmooth convex optimization.

- ◇ A (δ, L)-oracle can certify that an approximate solution has accuracy δ. Indeed, assuming $g_{\delta,L}(y)$ satisfies $\langle g_{\delta,L}(y), x - y\rangle \geq 0$ for all $x \in Q$, we have that $f_{\delta,L}(y) \leq f(x^*) = f^*$ and therefore, using (4.3), we have $f(y) \leq f^* + \delta$.

- ◇ If f admits a (δ, L)-oracle, then cf admits a $(c\delta, cL)$-oracle for any value of the constant $c > 0$. If f_i admits a (δ_i, L_i)-oracle, $i = 1, 2$, then $f_1 + f_2$ admits a $(\delta_1 + \delta_2, L_1 + L_2)$-oracle.

- ◇ When $Q = E$, the difference between $g_{\delta,L}$ and any subgradient $g_y \in \partial f(y)$ is bounded as follows

$$\|g_y - g_{\delta,L}(y)\|_E^* \quad \leq \quad [2\delta L]^{\frac{1}{2}}. \tag{4.5}$$

Indeed, for any $x \in Q$ we have $f(x) \geq f(y) + \langle g_y, x - y\rangle \geq f_{\delta,L}(y) + \langle g_y, x - y\rangle$. Subtracting this inequality from the second part of (4.2), we get that

$$\langle g_y - g_{\delta,L}(y), x - y\rangle \leq \frac{L}{2}\|x - y\|_E^2 + \delta$$

holds for all $x \in Q$. If $z \in E$ is such that $\|g_y - g_{\delta,L}(y)\|_E^* = |\langle g_y - g_{\delta,L}(y), z\rangle|$ and $\|z\|_E = 1$ and if we choose $x \in Q$ such that $x - y = t$ sign$(\langle g_y - g_{\delta,L}, z\rangle)z$ with $t > 0$, we obtain

$$t\|g_y - g_{\delta,L}(y)\|_E^* \leq \frac{L}{2}t^2 + \delta \Leftrightarrow \|g_y - g_{\delta,L}(y)\|_E^* \leq \frac{L}{2}t + \frac{\delta}{t}\,. \tag{4.6}$$

This upper bound attains its minimum $[2\delta L]^{\frac{1}{2}}$ when $t = [\frac{2\delta}{L}]^{\frac{1}{2}}$. In particular, when $Q = E$, parameter t is free to take any real value, and we obtain inequality (4.5). For constrained problems, a similar bound can be obtained in terms of the distance $d(y, \partial Q)$ between y and the boundary of Q: letting

$$d(y, \partial Q) = \max\{r | \|x - y\|_E \leq r \Rightarrow x \in Q\}$$

we have that (4.6) holds for all t such that $0 < t \leq d(y, \partial Q)$, so that

$$\|g_y - g_{\delta,L}(y)\|_E^* \leq \begin{cases} \frac{L}{2}d(y, \partial Q) + \frac{\delta}{d(y,\partial Q)} & \text{when } 0 < d(y, \partial Q) \leq [\frac{2\delta}{L}]^{\frac{1}{2}}\,, \\ [2\delta L]^{\frac{1}{2}} & \text{when } d(y, \partial Q) \geq [\frac{2\delta}{L}]^{\frac{1}{2}}\,. \end{cases}$$

◇ When E is endowed with a Euclidean norm (2.2), i.e. with $\|x\|_2^2 = \langle Bx, x\rangle$, the distance between exact and inexact gradient mappings can be bounded by the same quantities as the distance between exact and inexact
(sub)gradients. Recall that for any $\gamma > 0$, $g \in E^*$ and $y \in E$, the gradient mapping $M_\gamma(y, g)$, which replaces the gradient for constrained problems, is defined by

$$T_\gamma(y,g) = \arg\min_{x\in Q}\{\langle g, x-y\rangle + \frac{\gamma}{2}\|x-y\|_E^2\} \quad (4.7)$$

$$M_\gamma(y,g) = \gamma(y - T_\gamma(y,g)). \quad (4.8)$$

If f is subdifferentiable at point y, the exact gradient mapping for any subgradient $g_y \in \partial f(y)$ is equal to $M_\gamma(y, g_y)$. Similarly, if an inexact (δ, L) oracle returns $(f_{\delta,L}(y), g_{\delta,L}(y))$ for point y, we call $M_\gamma(y, g_{\delta,L}(y))$ the inexact gradient mapping. We are going to prove that the following holds

$$\|M_\gamma(y,g_y) - M_\gamma(y, g_{\delta,L}(y))\|_2 \le \|g_y - g_{\delta,L}(y)\|_2^* \quad (4.9)$$

(recall that the dual norm $\|x\|_2^*$ is defined in the Euclidean case by (2.3)).
First-order optimality conditions for (4.7) can be written as

$$\langle g + \gamma B(T_\gamma(y,g) - y), x - T_\gamma(y,g)\rangle \ge 0 \quad \forall x \in Q. \quad (4.10)$$

Applying those to $T_\gamma(y, g_y)$ and $T_\gamma(y, g_{\delta,L}(y))$ leads to

$$\begin{aligned}\langle g_y - BM_\gamma(y,g_y), x - T_\gamma(y,g_y)\rangle &\ge 0 \qquad \forall x \in Q\\ \langle g_{\delta,L}(y) - BM_\gamma(y,g_{\delta,L}(y)), x - T_\gamma(y,g_{\delta,L}(y))\rangle &\ge 0 \qquad \forall x \in Q\end{aligned}$$

and specializing respectively to $x = T_\gamma(y, g_{\delta,L}(y))$ and $x = T_\gamma(y, g_y)$ gives

$$\begin{aligned}\langle g_y - BM_\gamma(y,g_y), T_\gamma(y,g_{\delta,L}(y)) - T_\gamma(y,g_y)\rangle &\ge 0\\ \langle g_{\delta,L}(y) - BM_\gamma(y,g_{\delta,L}(y)), T_\gamma(y,g_y) - T_\gamma(y,g_{\delta,L}(y))\rangle &\ge 0\,.\end{aligned}$$

Using now (4.8) in the inner products, multiplying by γ and summing, we obtain

$$\langle g_y - BM_\gamma(y,g_y) - g_{\delta,L}(y) + BM_\gamma(y,g_{\delta,L}(y)), M_\gamma(y,g_y) - M_\gamma(y,g_{\delta,L}(y))\rangle \ge 0$$

which gives

$$\langle g_y - g_{\delta,L}(y), M_\gamma(y,g_y) - M_\gamma(y,g_{\delta,L}(y))\rangle \ge \|M_\gamma(y,g_y) - M_\gamma(y,g_{\delta,L}(y))\|_2^2\,,$$

from which the desired inequality (4.9) follows by Cauchy-Schwartz.

Characterizing the class of functions that can be endowed with a (δ, L)-oracle is an interesting open question. We provide below some necessary conditions in the simple case where $Q = E$ and E is endowed with an Euclidean norm (2.2). First of all, we establish the following inequality:

Theorem 4.1. *If f is equipped with a (δ, L)-oracle, we have*

$$\frac{1}{2L}(\|g_{\delta,L}(x) - g_{\delta,L}(y)\|_2^*)^2 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta \quad \forall x, y \in E.$$

Proof. Let us fix first $x \in E$ and $y \in E$, and assume a (δ, L)-oracle for function f returns $(f_{\delta,L}(z), g_{\delta,L}(z))$ for any $z \in E$. Introduce auxiliary functions $h(z) = -\langle g_{\delta,L}(x), z\rangle$ and $F(z) = f(z) + h(z)$. It is easy to check that $(-\langle g_{\delta,L}(x), z\rangle, -g_{\delta,L}(x))$ is a $(0,0)$-oracle for $h(z)$, and hence $(F_{\delta,L}(z), G_{\delta,L}(z)) = (f_{\delta,L}(z) - \langle g_{\delta,L}(x), z\rangle, g_{\delta,L}(z) - g_{\delta,L}(x))$ is a (δ, L)-oracle for $F(z)$.
The lower inequality valid for this oracle implies $F_{\delta,L}(w) + \langle G_{\delta,L}(w), z - w\rangle \leq F(z)$ for all $w, z \in E$. Applying this to $z = y - \frac{1}{L}B^{-1}G_{\delta,L}(y)$ and $w = x$, and noticing $G_{\delta,L}(x) = 0$, we find $F_{\delta,L}(x) \leq F\left(y - \frac{1}{L}B^{-1}G_{\delta,L}(y)\right)$. We now use the upper equality for the oracle $F(z) \leq F_{\delta,L}(w) + \langle G_{\delta,L}(w), z - w\rangle + \frac{L}{2}\|z - w\|_2^2 + \delta$ for all $w, z \in E$. Applying it to $z = y - \frac{1}{L}B^{-1}G_{\delta,L}(y)$ and $w = y$ we derive

$$\begin{aligned}
F_{\delta,L}(x) &\leq F\left(y - \frac{1}{L}B^{-1}G_{\delta,L}(y)\right) \\
&\leq F_{\delta,L}(y) + \langle G_{\delta,L}(y), -\frac{1}{L}B^{-1}G_{\delta,L}(y)\rangle + \frac{L}{2}\left\|\frac{1}{L}B^{-1}G_{\delta,L}(y)\right\|_2^2 + \delta \\
&= F_{\delta,L}(y) + \langle G_{\delta,L}(y), -\frac{1}{L}B^{-1}G_{\delta,L}(y)\rangle + \frac{1}{2L}\left(\|G_{\delta,L}(y)\|_2^*\right)^2 + \delta \\
&= F_{\delta,L}(y) - \frac{1}{2L}\left(\|G_{\delta,L}(y)\|_2^*\right)^2 + \delta
\end{aligned}$$

which allows us to obtain

$$\begin{aligned}
\frac{1}{2L}\left(\|g_{\delta,L}(y) - g_{\delta,L}(x)\|_2^*\right)^2 &\leq f_{\delta,L}(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta \\
&\leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta.
\end{aligned}$$

□

As a Corollary, we have:

Corollary 4.2. *If f is equipped with a (δ, L)-oracle, then we have for all $x, y \in E$*

$$\|g_{\delta,L}(x) - g_{\delta,L}(y)\|_2^* \leq \sqrt{L^2\|x - y\|_2^2 + 4L\delta}$$

and for any $g_x \in \partial f(x)$ and any $g_y \in \partial f(y)$

$$\|g_x - g_y\|_2^* \leq (2\sqrt{2} + 2)\sqrt{L\delta} + L\|x - y\|_2.$$

Proof. Our first claim directly follows from the previous theorem:

$$\begin{aligned}\frac{1}{2L}\left(\|g_{\delta,L}(x)-g_{\delta,L}(y)\|_2^*\right)^2 &\leq f(y)-f_{\delta,L}(x)-\langle g_{\delta,L}(x),y-x\rangle+\delta\\ &\overset{(4.2)}{\leq} \frac{L}{2}\|x-y\|_2^2+2\delta.\end{aligned}$$

Furthermore, for any $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$, we have by (4.5):

$$\|g_x - g_{\delta,L}(x)\|_2^* \leq \sqrt{2\delta L},$$

$$\|g_y - g_{\delta,L}(y)\|_2^* \leq \sqrt{2\delta L},$$

and therefore

$$\begin{aligned}\|g_x-g_y\|_2^* &\leq \|g_x-g_{\delta,L}(x)\|_2^*+\|g_{\delta,L}(x)-g_{\delta,L}(y)\|_2^*+\|g_{\delta,L}(y)-g_y\|_2^*\\ &\leq 2\sqrt{2\delta L}+\sqrt{L^2\|x-y\|_2^2+4L\delta}\\ &\leq (2\sqrt{2}+2)\sqrt{\delta L}+L\|x-y\|_2.\end{aligned}$$

where we used inequality $\sqrt{a^2+b^2} \leq |a|+|b|$ inequality on the last step. □

We conclude that the variation of subgradients of f is locally bounded:

$$\|g_x - g_y\|_2^* \leq (2\sqrt{2}+2)\sqrt{L\delta}+LR \quad \forall\, x,y \text{ s.t. } \|x-y\|_2 \leq R. \tag{4.11}$$

Note however that this property is true for any subdifferentiable convex function defined on the whole space E.

Assume now that, for a given function f, we can choose the oracle accuracy arbitrarily. This means that for any $\delta > 0$, there exists a constant $L(\delta)$ such that a $(\delta, L(\delta))$-oracle is available for f. We now apply inequality (4.11) to the following two situations:

◇
$$\lim_{\delta\to 0} L(\delta) = \bar{L} < +\infty$$
In this case we have $\|g_x - g_y\|_2^* \leq \bar{L}\|x-y\|_2$, so that f must be a smooth convex function with a Lipschitz-continuous gradient.

◇
$$\lim_{\delta\to\infty} L(\delta) = 0 \text{ and } \lim_{\delta\to\infty} L(\delta)\delta = \bar{C} < +\infty,$$
which is the case for example when $L(\delta) = \frac{\bar{C}}{\delta}$. We have then $\|g_x - g_y\|_2^* \leq (2\sqrt{2}+2)\sqrt{\bar{C}}$ so that f must be a convex function with bounded variation of subgradients.

Note that unless f is linear, $L(\delta)$ cannot decrease faster than the inverse of the oracle accuracy δ. Indeed we would have in that case $\bar{C} = 0$ and $\|g_x - g_y\|_2^* \leq 0$.

4.1.3 Examples

To conlude this section, we consider five simple examples of inexact oracle. More sophisticated examples will be given in Section 4.2.

a. Computations at shifted points. Let function $f \in F_M^{1,1}(Q)$ be endowed with an oracle providing at each point $y \in Q$ the exact values of function and gradient, albeit computed at a shifted point $\hat{y}$ different from y. Let us show that such an oracle can be converted into a (δ, L)-oracle with

$$\delta = M \|y - \hat{y}\|_E^2, \quad L = 2M.$$

Convexity of f implies the following inequality for any $x \in Q$

$$\begin{aligned} f(x) &\geq f(\hat{y}) + \langle \nabla f(\hat{y}), x - \hat{y} \rangle \\ &= f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \langle \nabla f(\hat{y}), x - y \rangle. \end{aligned}$$

Therefore, to satisfy the first inequality in (4.2) we can choose $f_{\delta,L}(y) \stackrel{\text{def}}{=} f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle$, and $g_{\delta,L}(y) \stackrel{\text{def}}{=} \nabla f(\hat{y})$. In order to prove the second inequality in (4.2), note that we have for all $x \in Q$

$$\begin{aligned} f(x) &\stackrel{(4.1)}{\leq} f(\hat{y}) + \langle \nabla f(\hat{y}), x - \hat{y} \rangle + \tfrac{M}{2} \|x - \hat{y}\|_E^2 \\ &= f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \langle \nabla f(\hat{y}), x - y \rangle + \tfrac{M}{2} \|x - \hat{y}\|_E^2. \end{aligned}$$

Since $\|\cdot\|_E^2$ is a convex function, we have

$$\begin{aligned} \|x - \hat{y}\|_E^2 &= \left\| \frac{1}{2}\big(2(x - y)\big) + \frac{1}{2}\big(2(y - \hat{y})\big) \right\|_E^2 \\ &\leq \frac{1}{2} \|2(x - y)\|_E^2 + \frac{1}{2} \|2(y - \hat{y})\|_E^2 = 2\|y - \hat{y}\|_E^2 + 2\|x - y\|_E^2. \end{aligned}$$

Therefore,

$$f(x) \leq f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y \rangle + M \|x - y\|_E^2 + M \|y - \hat{y}\|_E^2.$$

We can therefore choose $L = 2M$ and $\delta = M\|y - \hat{y}\|_E^2$ to satisfy the (δ, L)-oracle definition.

b. Functions approximated by a smooth function When a function f can be well approximated by a smooth convex function $\bar{f}$, in the sense that their difference is bounded, the exact values of $\bar{f}$ and its gradient provide an inexact oracle for f. Indeed, assume that there exists a smooth convex function $\bar{f} \in F_L^{1,1}(Q)$ such that $\bar{f}$ is a δ-lower approximation of f on all Q, i.e.

$$0 \leq f(y) - \bar{f}(y) \leq \delta \quad \forall y \in Q.$$

We conclude that

$$f(x) \geq \bar{f}(x) \geq \bar{f}(y) + \langle \nabla \bar{f}(y), x - y \rangle \quad \forall x, y \in Q,$$

(using convexity of $\bar{f}$), and

$$f(x) \leq \bar{f}(x) + \delta \leq \bar{f}(y) + \langle \nabla \bar{f}(y), x - y \rangle + \frac{L}{2} \|x - y\|_E^2 + \delta \quad \forall x, y \in Q.$$

(using Lipschitz continuity of $\nabla \bar{f}$), which shows that $(\bar{f}(y), \nabla \bar{f}(y))$ is a (δ, L)-oracle for f.

One might wonder whether all inexact oracles can be obtained in that fashion, i.e. whether any inexact oracle can be seen as an exact oracle for a smooth approximation $\bar{f}$. It turns out that is not the case: indeed, as we have seen earlier, when f has subgradients with bounded variation, its exact function values and subgradients can be seen as a (δ, L)-oracle (for arbitrary value of δ). Clearly, such an oracle cannot be at the same time equal to the exact function values and gradients of any smooth function $\bar{f}$.

Finally, note that the above result can be readily extended to the case when the δ-lower approximation $\bar{f}$ is not necessarily smooth but is equipped with an inexact (δ', L) oracle: we can then show that the inexact oracle of $\bar{f}$ also constitutes an inexact $(\delta + \delta', L)$ oracle for f.

c. Convex problems with weaker level of smoothness. Let us show that the notion of (δ, L)-oracle can be useful for solving problems with exact first-order information but with a lower level of smoothness. Let function f be convex and subdifferentiable on Q. For each $y \in Q$, denote by $g(y)$ an arbitrary element of the subdifferential $\partial f(y)$. Assume that f satisfies the following Hölder condition:

$$\|g(x) - g(y)\|_E^* \quad \leq \quad L_\nu \|x - y\|_E^\nu, \quad \forall x, y \in Q, \tag{4.12}$$

where $\nu \in [0, 1]$, and $L_\nu < +\infty$. This condition leads to the following inequality:

$$f(x) \quad \leq \quad f(y) + \langle g(y), x - y \rangle + \tfrac{L_\nu}{1+\nu} \|x - y\|_E^{1+\nu}, \quad \forall x, y \in Q. \tag{4.13}$$

Denote the class of such functions by $F_{L_\nu}^{1,\nu}(Q)$. When $\nu = 1$, we get functions with Lipschitz-continuous gradient. For $\nu < 1$, we get a lower level of smoothness. In particular, when $\nu = 0$, we obtain functions whose subgradients have bounded variation. Clearly, the latter class includes functions whose subgradients are uniformly bounded by M (just take $L_0 = 2M$).

Let us fix $\nu \in [0, 1)$ and an arbitrary $\delta > 0$. We are going to find a constant $A(\delta, \nu)$ such that for any function $f \in F_{L_\nu}^{1,\nu}(Q)$ we have

$$f(x) - f(y) - \langle g(y), x - y\rangle \quad \leq \quad \tfrac{A(\delta,\nu)}{2} \|x - y\|_E^2 + \delta, \quad \forall x, y \in Q. \tag{4.14}$$

Comparing (4.13) and (4.14), we need choose $A(\delta, \nu)$ such that

$$\frac{L_\nu}{1+\nu} \|x - y\|_E^{1+\nu} \ \leq \ \frac{A(\delta, \nu)}{2} \|x - y\|_E^2 + \delta \ .$$

Since $t = \|x - y\|_E^2$ can take any nonnegative value, we may choose

$$A(\delta, \nu) \quad = \quad 2 \max_{t \geq 0} \left\{ \tfrac{L_\nu}{1+\nu} t^{-1+\nu} - \delta t^{-2} \right\} = L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}$$

(the latter expression is obtained after straightforward computations, the optimal value of t in the maximization being $t_* = \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu} \right]^{-\frac{1}{1+\nu}}$). This means that the exact first-order information $(f(y), g(y))$ also constitutes an inexact $(\delta, A(\delta, \nu))$-oracle. We will therefore be able to apply the methods from Sections 4.3 and 4.4, initially devoted for smooth problems, to the minimization of the non- or weakly smooth objective f.

For example, for functions with bounded variation of subgradients ($\nu = 0$) we have

$$A(\delta, 0) \quad = \quad \tfrac{L_0^2}{2\delta}. \tag{4.15}$$

so that a $(\delta, \frac{L_0^2}{2\delta})$-oracle is available for all values of $\delta > 0$.

Note that parameter δ does not represent an actual accuracy: it can be chosen arbitrarily, independently of the answer of the oracle. In particular, δ can be chosen as small as we want, at the price of a larger value for Lipschitz constant L of the (δ, L)-oracle, which grows as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$.

The application of first-order methods of smooth convex optimization to nonsmooth or weakly smooth functions will be be further investigated in Section 4.7, where parameter δ will be tuned to minimize the overall complexity of the considered methods.

Remark 4.1. This analysis can easily be extended to the case where δ-subgradients with bounded variation are used instead of exact subgradients. We obtain in this case a $(2\delta, A(\delta,\nu))$-oracle.

d. Function smoothed by local averaging. Another way to apply first-order methods of smooth convex optimization to a nonsmooth function consists in smoothing the function by averaging first-order information. Assume that E is endowed with an Euclidean norm and consider a nonsmooth convex function $f \in F_M^{1,0}(E)$. Let $r > 0, y \in E$, and define

$$\begin{aligned} f_r(y) &= \frac{1}{V_r}\int_{\|z-y\|_2 \le r} f(z)\,\mathrm{d}z \\ \nabla f_r(y) = g_r(y) &= \frac{1}{V_r}\int_{\|z-y\|_2 \le r} g(z)\,\mathrm{d}z \end{aligned}$$

where V_r denotes the volume of a Euclidean ball with radius r, and $\{g(z) : \|z-y\|_2 \le r\}$ is a measurable selection of subgradients of f in this ball. As f is convex and Lipschitz-continuous with constant M we have

$$0 \le f(x) - f(z) - \langle g(z), x - z\rangle \le M\|x-z\|_2 \quad \forall x, z \in E$$

and therefore

$$\begin{aligned} f(x) &\ge f(z) + \langle g(z), x-y\rangle + \langle g(z), y-z\rangle \quad \forall x,y,z \in E \\ f(x) &\le f(z) + \langle g(z), x-y\rangle + \langle g(z), y-z\rangle + M\|x-z\|_2 \quad \forall x,y,z \in E. \end{aligned}$$

Averaging now these two inequalities with respect to z over the ball $\{z : \|z-y\|_2 \le r\}$, we obtain for all $x, y \in Z$

$$\begin{aligned} f(x) &\ge f_r(y) + \langle g_r(y), x-y\rangle - Mr \\ f(x) &\le f_r(y) + \langle g_r(y), x-y\rangle + Mr + \frac{M}{V_r}\int_{\|z-y\|_2\le r}\|x-z\|_2\,\mathrm{d}z \end{aligned}$$

(where we used that $|\langle g(z), y-z\rangle| \le \|g(z)\|_2^*\|y-z\|_2 \le Mr$). Furthermore, we have

$$\|x-z\|_2 \overset{(4.12)}{\le} \sqrt{2\|x-y\|_2^2 + 2\|z-y\|_2^2} \le \frac{2\|x-y\|_2^2 + 2\|z-y\|_2^2}{2r} + \frac{r}{2}$$

(where the second inequality comes from the arithmetic-geometric inequality), and therefore

$$\begin{aligned} f(x) &\le f_r(y) + \langle g_r(y), x-y\rangle + \frac{3}{2}Mr + M\frac{\|x-y\|_2^2}{r} + \frac{M}{V_r}\int_{\|z-y\|_2\le r}\|z-y\|_2\,\mathrm{d}z \\ &\le f_r(y) + \langle g_r(y), x-y\rangle + \frac{5}{2}Mr + M\frac{\|x-y\|_2^2}{r}. \end{aligned}$$

Finally, choosing $f_{\delta,L}(y) = f_r(y) - Mr$, $g_{\delta,L}(y) = g_r(y)$, $\delta = \frac{7Mr}{2}$ and $L = \frac{2M}{r}$, we obtain a $(\delta, L) = (\frac{7Mr}{2}, \frac{2M}{r}) = (\delta, \frac{7M^2}{\delta})$-oracle. Note that the dependence of L in M and δ is similar to that of the previous example, where subgradients are used directly instead of being averaged.

e. Approximate function value and approximate gradient Let function $f \in F_M^{1,1}(Q)$ be endowed with an oracle that provides us at each point $y \in Q$ with an approximate function value $|f(y) - \tilde{f}_y| \leq \Delta_1$ and an approximate gradient $\left\|\nabla f(y) - \tilde{\nabla} f_y\right\|_E^* \leq \Delta_2$.

When the set Q is bounded (with diameter D), this very natural definition of approximate first-order information is a particular case of (δ, L) oracle: $(f_{\delta,L}(y) = \tilde{f}_y - \Delta_1 - \Delta_2 D$, $g_{\delta,L}(y) = \tilde{\nabla} f_y)$ is a (δ, L) oracle with $\delta = 2\Delta_1 + 2\Delta_2 D$ and $L = M$.

Indeed, we have:

$$\begin{aligned} f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle \\ &\geq \tilde{f}_y - \Delta_1 + \langle \tilde{\nabla} f_y, x - y \rangle - \Delta_2 D = f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y \rangle. \end{aligned}$$

and

$$\begin{aligned} f(x) &\leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|_E^2 \\ &\leq \tilde{f}_y + \Delta_1 + \langle \tilde{\nabla} f_y, x - y \rangle + \Delta_2 D + \frac{L}{2} \|x - y\|_E^2 \\ &= f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y \rangle + \frac{L}{2} \|x - y\|_E^2 + 2\Delta_1 + 2\Delta_2 D. \end{aligned}$$

Remark 4.2. The fact that an approximate gradient is a particular case of (δ, L)-oracle, is only true in the bounded case. In the unbounded case, it is easy to construct an example of approximate gradient which is not a (δ, L)-oracle. Indeed consider $f(x) = \frac{M}{2} x_{(1)}^2 + \frac{\Delta}{2} x_{(2)}$ and $Q = \mathbb{R}^2$ endowed with the classical Euclidean norm on $\mathbb{R}^2$. This function is clearly in $F_M^{1,1}(\mathbb{R}^2)$ and for all $y \in \mathbb{R}^2$, the approximate gradient $\tilde{\nabla} f_y = (My_{(1)}, -\frac{\Delta}{2})$ satisfies $\left\|\tilde{\nabla} f_y - \nabla f(y)\right\|_2 \leq \Delta$. However there exists no $\tilde{f}_y \in \mathbb{R}, L \geq 0$ and $\delta \geq 0$ such that $(\tilde{f}_y, \tilde{\nabla} f_y)$ is a (δ, L) oracle. Indeed, the condition

$$f(x) \geq \tilde{f}_y + \langle \tilde{\nabla} f_y, x - y \rangle, \quad \forall x \in \mathbb{R}$$

is equivalent with

$$\frac{M}{2}(y_{(1)} - x_{(1)})^2 \geq \Delta(y_{(2)} - x_{(2)}) + (\tilde{f}_y - f(y)), \quad \forall x \in \mathbb{R}.$$

Taking $y_{(1)} = x_{(1)}$, we obtain the condition

$$\Delta(x_{(2)} - y_{(2)}) \geq \tilde{f}_y - f(y), \quad \forall x_{(2)} \in \mathbb{R}$$

which is clearly impossible to satisfy when $\Delta \neq 0$.

Remark 4.3. We have seen in equation (4.5) that when $Q = E$, the reverse implication holds: a (δ, L)-oracle is always an approximate gradient.

4.2 (δ, L)-oracle for functions defined by an optimization subproblem

4.2.1 Accuracy measures for approximate solutions

In this section, we consider smooth convex optimization problems of the form (2.1) whose objective function $f \in F^{1,1}_{L(f)}(Q)$ is defined by another optimization problem:

$$f(x) \quad = \quad \max_{u \in U} \Psi(x, u), \tag{4.16}$$

where U is a convex set of a finite dimensional space F endowed with the norm $\|.\|_F$ and for any $x \in Q$ function $\Psi(x, \cdot)$ is smooth and (strongly) concave with concavity parameter $\kappa \geq 0$. Computation of f and its gradient requires the exact solution of this auxiliary problem. However, in practice, such a solution might often be impossible or too costly to compute, so that an approximate solution has to be used instead.

We will measure the accuracy of an approximate solution u_x for problem (4.16) in three different ways:

$$\begin{aligned} V_1(u_x) &= \max_{u \in U} \langle \nabla_2 \Psi(x, u_x), u - u_x \rangle, \\ V_2(u_x) &= \max_{u \in U} \left[\Psi(x, u) - \Psi(x, u_x) + \tfrac{\kappa}{2} \|u_x - u\|_F^2 \right], \\ V_3(u_x) &= \max_{u \in U} \left[\Psi(x, u) - \Psi(x, u_x) \right]. \end{aligned} \tag{4.17}$$

Since $\Psi(x, \cdot)$ is (strongly) concave, we have:

$$\Psi(x, u) \quad \leq \quad \Psi(x, u_x) + \langle \nabla_2 \Psi(x, u_x), u - u_x \rangle - \tfrac{\kappa}{2} \|u - u_x\|_F^2, \quad \forall\, u \in U.$$

Therefore our three measures are related by

$$V_3(u_x) \quad \leq \quad V_2(u_x) \; \leq \; V_1(u_x).$$

For a given level of accuracy $\delta > 0$, the condition $V_1(u_x) \leq \delta$ is the strongest, and condition $V_3(u_x) \leq \delta$ is the most relaxed.

We describe below three classes of max-type functions for which the approximate solution of subproblem (4.16), when satisfying one of the conditions $V_i(u_x) \leq \delta$, allows the construction of a (δ, L)-oracle.

Let us show first how to satisfy stopping criteria (4.17) in practice. The most common criterion is the third one. It amounts to estimating the optimality gap in the value of objective function. Many optimization methods offer direct control of this criterion. Other criteria might be more difficult to handle. Therefore, let us describe a "brute force" approach designed to satisfy the strongest V_1 criterion. We assume here that F is endowed with an Euclidean norm, that the set U is bounded with diameter D_U and that $\Psi(x, .)$ has a Lipschitz continuous gradient with constant L_U. Denote $h(u) = -\Psi(x, u)$ and $\overline{u}$ an arbitrary point in U. Then, after a gradient step from $\overline{u}$, we obtain a new point $\overline{v} = T_{L_U}(\overline{u}, \nabla h(\overline{u}))$ such that

$$\langle \nabla h(\overline{u}) + L_U(\overline{v} - \overline{u}), u - \overline{v} \rangle \geq 0, \quad \forall u \in U.$$

Therefore:

$$\begin{aligned} \langle \nabla h(\overline{v}), \overline{v} - u \rangle &= \langle \nabla h(\overline{v}) - \nabla h(\overline{u}), \overline{v} - u \rangle + \langle \nabla h(\overline{u}), \overline{v} - u \rangle \\ &\leq L_U \left\| \overline{v} - \overline{u} \right\| D_U + \langle L_U(\overline{v} - \overline{u}), u - \overline{v} \rangle \\ &\leq 2L_U \left\| \overline{v} - \overline{u} \right\| D_U \\ &= 2 \left\| M_{L_U}(\overline{u}, \nabla h(\overline{u})) \right\| D_U \end{aligned}$$

where $M_{L_U}(\overline{u}, \nabla h(\overline{u})) = L_U(\overline{u} - T_{L_U}(\overline{u}, \nabla h(\overline{u})))$ denotes the gradient mapping of h at $\overline{u}$.

We obtain that the quality of the point $\overline{v}$ for the criterion V_1 can be bounded by:

$$\begin{aligned} V_1(\overline{v}) &= \max_{u \in U} \langle \nabla_2 \Psi(x, \overline{v}), u - \overline{v} \rangle \\ &= \max_{u \in U} \langle \nabla h(\overline{v}), \overline{v} - u \rangle \leq 2 \left\| M_{L_U}(\overline{u}, \nabla h(\overline{u})) \right\| D_U. \end{aligned}$$

The upper-bound $V_1(\overline{v}) \leq 2 \left\| M_{L_U}(\overline{u}, \nabla h(\overline{u})) \right\| D_U$

1. is computable in practice, hence the quality of $\overline{v}$ for the criterion V_1 can be checked directly ;

2. is decreasing to zero for any convergent optimization scheme applied to

h. Indeed, we have (see Corollary 2.2.1 in [58])

$$\begin{aligned} \|M_{L_U}(\overline{u}, \nabla h(\overline{u}))\| &\leq \sqrt{2L_U(h(\overline{u}) - h(\overline{v}))} \\ &\leq \sqrt{2L_U(h(\overline{u}) - h^*)}. \end{aligned}$$

If we apply the fast gradient method to the function $h(u)$, generating the sequence of iterates u_k, and if we take one gradient step from u_k, obtaining v_k, then $V_1(v_k)$ goes to zero with the rate $O\left(\frac{1}{k}\right)$.

4.2.2 Functions obtained by smoothing techniques

Let U be a closed, convex set of a finite dimensional space F endowed with the norm $\|\cdot\|_F$, and

$$\Psi(x, u) = G(u) + \langle Au, x\rangle,$$

where $A : F \to E^*$ is a linear operator, and $G(u)$ is a differentiable, strongly concave function with concavity parameter $\kappa > 0$. Under these assumptions, optimization problem (4.16) has only one optimal solution u_x^*. Moreover, f is convex and smooth with Lipschitz-continuous gradient $\nabla f(x) = Au_x^*$. The corresponding Lipschitz-constant is equal to

$$L(f) \quad = \quad \tfrac{1}{\kappa}\|A\|^2_{F\to E^*} \tag{4.18}$$

where $\|A\|_{F\to E^*} = \max\{\|Au\|_{E^*} : \|u\|_F = 1\}$. The importance of this class of functions is justified by the smoothing approach for nonsmooth convex optimization (see subsection 2.5.3 in Chapter 2, Chapter 3 and [59, 60, 61, 23]).

Suppose that for all $y \in Q$ we can find a point $u_y \in U$ satisfying condition

$$V_3(u_y) \quad = \quad \Psi(y, u_y^*) - \Psi(y, u_y) \;\leq\; \tfrac{\delta}{2}. \tag{4.19}$$

Let us show that this allows us to construct an $(\delta, 2L(f))$-oracle. Indeed, since $\Psi(\cdot, u)$ is convex, for all $u \in U$, we have

$$\begin{aligned} f(x) = \Psi(x, u_x^*) &\geq \Psi(x, u_y) \\ &\geq \Psi(y, u_y) + \langle \nabla_1\Psi(y, u_y), x - y\rangle \\ &= f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y\rangle, \end{aligned} \tag{4.20}$$

where $f_{\delta,L}(y) \stackrel{\text{def}}{=} \Psi(y, u_y)$, $g_{\delta,L}(y) \stackrel{\text{def}}{=} \nabla_1\Psi(y, u_y) = Au_y$, and L will be specified later. Further, note that

$$\langle \nabla_1\Psi(y, u_y^*), x - y\rangle \quad = \quad \langle g_{\delta,L}(y), x - y\rangle + \langle A(u_y^* - u_y), x - y\rangle. \tag{4.21}$$

Since f has Lipschitz-continuous gradient, we have:

$$\begin{aligned} f(x) &\leq f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{L(f)}{2} \|x - y\|_E^2 \\ &= f(y) + \langle \nabla \Psi_1(y, u_y^*), x - y \rangle + \tfrac{L(f)}{2} \|x - y\|_E^2 \\ &\overset{(4.21)}{=} f(y) + \langle g_{\delta,L}(y), x - y \rangle + \tfrac{L(f)}{2} \|x - y\|_E^2 + \langle A(u_y^* - u_y), x - y \rangle. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \langle A(u_y^* - u_y), x - y \rangle &\leq \left\| u_y^* - u_y \right\|_F \left\| A^T(x - y) \right\|_E^* \\ &\overset{(4.18)}{\leq} \tfrac{\kappa}{2} \left\| u_y^* - u_y \right\|_F^2 + \tfrac{L(f)}{2} \|x - y\|_E^2 . \end{aligned}$$

(where we used $\sqrt{ab} \leq \frac{a+b}{2}$). Therefore,

$$f(x) \leq f(y) + \langle g_{\delta,L}(y), x - y \rangle + L(f) \|x - y\|_E^2 + \tfrac{\kappa}{2} \left\| u_y^* - u_y \right\|_F^2 .$$

Since Ψ is strongly concave, $\frac{\kappa}{2} \left\| u_y - u_y^* \right\|_F^2 \leq \Psi(y, u_y^*) - \Psi(y, u_y)$. Thus,

$$f(x) \leq \Psi(y, u_y) + 2(\Psi(y, u_y^*) - \Psi(y, u_y)) + \langle g_{\delta,L}(y), x - y \rangle + L(f) \|x - y\|_E^2 .$$

In view of conditions (4.19) and (4.20), we have proved that the pair $(\Psi(y, u_y), Au_y)$, satisfying condition (4.19), corresponds to a (δ, L)-oracle with $L = 2L(f)$.

4.2.3 Moreau-Yosida regularization

In this section, we consider functions of the form

$$f(x) = \min_{u \in U} \left\{ \mathcal{L}(x, u) \overset{\text{def}}{=} h(u) + \tfrac{\kappa}{2} \|u - x\|_2^2 \right\}, \tag{4.22}$$

where h is a smooth convex function on a convex set $U \subset \mathbb{R}^n$ endowed with the usual Euclidean norm $\|x\|_2^2 = \langle x, x \rangle$. The function f is convex with Lipschitz-continuous gradient $\nabla f(x) = \kappa(x - u_x^*)$, where u_x^* denotes the unique optimal solution of the problem (4.22). The Lipschitz constant of the gradient is equal to κ.

Instead of solving exactly problem (4.22), we compute a feasible solution u_x satisfying

$$V_2(u_x) = \max_{u \in U} \left\{ \mathcal{L}(x, u_x) - \mathcal{L}(x, u) + \tfrac{\kappa}{2} \|u - u_x\|_2^2 \right\} \leq \delta. \tag{4.23}$$

(Since $\mathcal{L}$ is convex in u, we inverted the sign in the definition of V_2 in (4.17).) Let us show that for all $x \in Q$ the objects

$$\begin{aligned} f_{\delta,L}(x) &= \mathcal{L}(x,u_x) - \delta \;=\; h(u_x) + \tfrac{\kappa}{2}\|u_x - x\|_2^2 - \delta, \\ g_{\delta,L}(x) &= \nabla_1 \mathcal{L}(x,u_x) \;=\; \kappa(x - u_x) \end{aligned} \tag{4.24}$$

correspond to an answer of an (δ, L)-oracle with $L = \kappa$. Indeed,

$$\begin{aligned} f(x) &= \mathcal{L}(x,u_x^*) \;\geq \mathcal{L}(y,u_x^*) + \kappa\langle y - u_x^*, x - y\rangle \\ &\geq \mathcal{L}(y,u_x^*) + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - x - y\rangle \\ &\overset{(4.23)}{\geq} \mathcal{L}(y,u_y) + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - x - y\rangle \\ &= \mathcal{L}(y,u_y) + \kappa\langle y - u_y, x - y\rangle + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta \\ &\quad + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - 2u_y + y - x\rangle \\ &= \mathcal{L}(y,u_y) + \kappa\langle y - u_y, x - y\rangle - \delta \\ &\quad + \tfrac{\kappa}{2}\left(\|u_x^* - u_y\|_2^2 + \|y - x\|_2^2 + 2\langle y - x, u_x^* - u_y\rangle\right) \\ &\geq \mathcal{L}(y,u_y) + \kappa\langle y - u_y, x - y\rangle - \delta. \end{aligned}$$

Thus, we satisfy the first inequality in (4.2) with the values defined by (4.24). Further, for all $x, y \in Q$ we have

$$\begin{aligned} f(x) &= h(u_x^*) + \tfrac{\kappa}{2}\|u_x^* - x\|_2^2 \\ &\leq h(u_y) + \tfrac{\kappa}{2}\|u_y - x\|_2^2 \\ &= h(u_y) + \tfrac{\kappa}{2}\|u_y - y\|_2^2 + \tfrac{\kappa}{2}\langle x - y, x + y - 2u_y\rangle \\ &= \mathcal{L}(y,u_y) + \kappa\langle y - u_y, x - y\rangle + \tfrac{\kappa}{2}\|y - x\|_2^2. \end{aligned}$$

Thus, in view of definition (4.24), we have proved the second inequality in (4.2) with $L = \kappa$.

4.2.4 Functions defined by Augmented Lagrangians

Consider the following convex problem:

$$\max_{u \in U} \{h(u) : \; Au = 0\}, \tag{4.25}$$

where h is a smooth concave function on the convex set $U \subset F$, F is a finite-dimensional space, and $A : F \to E^*$ is a linear operator. Let E be endowed with the Euclidean norm $\|.\|_2$.

In the Augmented Lagrangian approach, we need to solve the dual problem

$$\min_{x \in E} f(x), \tag{4.26}$$

where

$$f(x) \stackrel{\text{def}}{=} \max_{u \in U} \left[\Psi(x,u) \stackrel{\text{def}}{=} h(u) + \langle Au, x \rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2 \right]. \tag{4.27}$$

It is well-known that f is a smooth convex function with Lipschitz-continuous gradient

$$\nabla f(x) = Au_x^*,$$

where u_x^* denotes any optimal solution of the optimization problem (4.27). The Lipschitz constant of the gradient is equal to $\frac{1}{\kappa}$.

Assume that, instead of solving (4.26) exactly, we compute an approximate solution $u_x \in U$ such that

$$\begin{aligned} V_1(u_x) &= \max_{u \in U} \langle \nabla_2 \Psi(x, u_x), u - u_x \rangle \\ &= \max_{u \in U} \langle \nabla h(u_x) + A^T x - \kappa A^T B^{-1} A u_x, u - u_x \rangle \ \leq \ \delta. \end{aligned} \tag{4.28}$$

Let us show that the objects

$$f_{\delta,L}(x) \ = \ \Psi(x, u_x), \quad g_{\delta,L}(x) \ = \ \nabla_1 \Psi(x, u_x) \ = \ Au_x \tag{4.29}$$

correspond to a (δ, L)-oracle with $L = \frac{1}{\kappa}$. Indeed, for all $x, y \in E$ we have

$$\begin{aligned} f(x) &= \max_{u \in U} \left\{ h(u) + \langle Au, x \rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2 \right\} \\ &\geq h(u_y) + \langle Au_y, x \rangle - \tfrac{\kappa}{2}(\|Au_y\|_2^*)^2 \ = \ \Psi(y, u_y) + \langle Au_y, x - y \rangle. \end{aligned}$$

Thus, in view of definition (4.29), the left inequality in (4.2) is proved. Further,

$$
\begin{aligned}
f(x) \quad &\leq \quad \max_{u \in U} \{h(u_y) + \langle \nabla h(u_y), u - u_y \rangle + \langle Au, x \rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\} \\
&\overset{(4.28)}{\leq} \quad \max_{u \in U} \{h(u_y) - \langle A^T y - \kappa A^T B^{-1} A u_y, u - u_y \rangle \\
&\qquad\qquad + \langle Au, x \rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\} + \delta \\
&= \quad \Psi(y, u_y) + \langle Au_y, x - y \rangle \\
&\qquad\qquad + \max_{u \in U} \left\{ \langle A(u - u_y), x - y \rangle - \tfrac{\kappa}{2}(\|A(u - u_y)\|_2^*)^2 \right\} + \delta \\
&= \quad \Psi(y, u_y) + \langle Au_y, x - y \rangle + \tfrac{1}{2\kappa} \|x - y\|_2^2 + \delta
\end{aligned}
$$

Thus, in view of (4.29), we have proved the right inequality in (4.2) with $L = \frac{1}{\kappa}$.

4.3 Gradient Methods with (δ, L)-oracle

Consider the convex optimization problem $\min_{x \in Q} f(x)$, where f is endowed with a (δ, L)-oracle. In this section, for the simplicity of the analysis, we use the Euclidean setup.

4.3.1 Primal gradient method

The classical (primal) gradient method (see subsection 2.4.2) can be adapted in a straightforward manner to accept first-order information from an inexact oracle: it is enough to replace the true gradient by its approximate counterpart $g_{\delta,L}$. We obtain

Algorithm 13 Primal Gradient Method (PGM) with (δ, L)-oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$.
4: Compute $x_{k+1} = T_L(x_k, g_{\delta,L}(x_k))$.
5: **end for**

Lemma 4.3. *For $k \geq 1$, we have*

$$
\sum_{i=0}^{k-1} [f(x_{i+1}) - f(x^*)] \quad \leq \quad \tfrac{L}{2} \|x_0 - x^*\|_2^2 + k\delta. \tag{4.30}
$$

Proof. Denote $r_k = \|x_k - x^*\|_2^2$, $f_k = f_{\delta,L}(x_k)$, and $g_k = g_{\delta,L}(x_k)$. Then

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + 2\langle B(x_{k+1} - x_k), x_{k+1} - x^* \rangle - \|x_{k+1} - x_k\|_2^2 \\ &\overset{(4.10)}{\leq} r_k^2 + \tfrac{2}{L}\langle g_k, x^* - x_{k+1} \rangle - \|x_{k+1} - x_k\|_2^2 \\ &= r_k^2 + \tfrac{2}{L}\langle g_k, x^* - x_k \rangle - \tfrac{2}{L}[\langle g_k, x_{k+1} - x_k \rangle + \tfrac{L}{2}\|x_{k+1} - x_k\|_2^2] \\ &\overset{(4.2)}{\leq} r_k^2 + \tfrac{2}{L}[f(x^*) - f_k] - \tfrac{2}{L}[f(x_{k+1}) - f_k - \delta]. \end{aligned}$$

Summing up these inequalities for $i = 0, \ldots, k-1$, we obtain (4.30). □

When exact first-order information is used ($\delta = 0$, $L = L(f)$), it is well-known that sequence $\{f(x_i)\}_{i \geq 0}$ must be decreasing. This is no longer true when an inexact oracle is used. Therefore, let us define

$$y_k = \frac{\sum_{i=1}^k x_i}{k} \in Q.$$

Since f is convex, we obtain the following convergence rate

Theorem 4.4. *Assume that function f is endowed with a (δ, L)-oracle. Then the sequence $y_k = \frac{\sum_{i=1}^k x_i}{k}$ generated by the PGM satisfies*

$$f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k} + \delta \leq \frac{LR^2}{2k} + \delta. \tag{4.31}$$

Thus, there is no error accumulation, and the upper bound for the objective function accuracy decreases with k and asymptotically tends to δ. Hence, if an accuracy ϵ on the objective function is required (with $\epsilon > \delta$), $k = \frac{LR^2}{2(\epsilon - \delta)}$ iterations are sufficient. In particular, we see that PGM allows the oracle accuracy to be of the same order as the desired accuracy for the objective function.

Remark 4.4. The same convergence result can be obtained for the non-Euclidean PGM with $\frac{\|x_0 - x^*\|_2^2}{2}$ replaced by $V(x^*, x_0)$. See Theorems 7.1 and 7.2 in Chapter 7 with $\phi = f$, $\sigma = 0$ and $\gamma_i = \frac{1}{L}$ for all $i \geq 0$.

4.3.2 Dual gradient method

When used with a (δ, L) oracle, the Euclidean DGM, developed in [62] and described in the subsection 2.4.4 of this thesis, becomes

Algorithm 14 Dual Gradient Method (DGM) with (δ, L)-oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$.
4: Compute $w_k = T_L(x_k, g_{\delta,L}(x_k))$
5: Compute $x_{k+1} = \arg\min_{x \in Q} \left[\sum_{i=0}^{k} \langle g_{\delta,L}(x_i), x - x_i \rangle + \frac{L}{2}\|x - x_0\|_2^2 \right]$.
6: **end for**

Lemma 4.5. *For any $k \geq 0$ we have*

$$\sum_{i=0}^{k} [f(w_i) - f^*] \quad \leq \quad \tfrac{L}{2}\|x_0 - x^*\|_2^2 + (k+1)\delta. \tag{4.32}$$

Proof. For $k \geq 0$, denote $f_k = f_{\delta,L}(x_k)$, $g_k = g_{\delta,L}(x_k)$, and

$$\psi_k(x) \quad = \quad \sum_{i=0}^{k} [f_i + \langle g_i, x - x_i \rangle] + \tfrac{L}{2}\|x - x_0\|_2^2, \quad \psi_k^* \; = \; \min_{x \in Q} \psi_k(x).$$

In view of the first inequality in (4.2), we have for all $x \in Q$

$$\psi_k^* \quad \leq \quad \psi_k(x) \; \leq \; \textstyle\sum_{i=0}^{k} f(x) + \frac{L}{2}\,\|x - x_0\|_2^2 \,. \tag{4.33}$$

Let us prove that $\psi_k^* \geq \sum\limits_{i=0}^{k} f(w_i) - (k+1)\delta$. Indeed, this inequality is valid for $k = 0$:

$$f(w_0) \quad \overset{(4.2)}{\leq} \quad f_0 + \langle g_0, w_0 - x_0 \rangle + \tfrac{L_0}{2}\,\|w_0 - x_0\|_2^2 + \delta \; = \psi_0^* + \delta.$$

Assume it is valid for some $k \geq 1$. Since $\Psi_k(x)$ is strongly convex with parameter L, we have:

$$\psi_k(x) \quad \geq \quad \psi_k^* + \tfrac{L}{2}\|x - x_{k+1}\|_2^2, \quad x \in Q$$

Therefore,

$$\begin{aligned} \psi_{k+1}^* \quad &= \quad \min_{x \in Q} \{\psi_k(x) + [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle]\} \\ &\geq \quad \psi_k^* + \min_{x \in Q} \left\{ f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \tfrac{L}{2}\|x - x_{k+1}\|_2^2 \right\} \\ &\overset{(4.2)}{\geq} \quad \psi_k^* + f(w_{k+1}) - \delta. \end{aligned}$$

Hence, using our inductive assumption, we have proved that $\psi_k^* \geq \sum\limits_{i=0}^{k} f(w_i) - (k+1)\delta$ for all $k \geq 0$. To conclude, we combine this fact with inequality (4.33) for $x = x^*$. □

Like in the exact case, we define the approximate solution as

$$y_k \quad = \quad \frac{\sum_{i=0}^{k} w_i}{k+1} \ \in \ Q,$$

and obtain the same kind of convergence rate:

Theorem 4.6. *Assume that f is endowed with a (δ, L)-oracle. Then the sequence $y_k = \frac{\sum_{i=0}^{k} w_i}{k+1}$ generated by the DGM satisfies*

$$f(y_k) - f^* \quad \leq \quad \frac{L\|x_0 - x^*\|_2^2}{2(k+1)} + \delta, \quad k \geq 0. \tag{4.34}$$

Remark 4.5. The same convergence result can be obtained for the non-Euclidean DGM with $\frac{\|x_0 - x^*\|_2^2}{2}$ by $d(x^*)$. See Lemma 7.3 and Theorem 7.4 in Chapter 7 with $\phi = f$, $\sigma = 0$, $\alpha_i = 1$ and $\beta_i = L$ for all $i \geq 0$.

Since we obtain the same convergence results for both primal and dual gradient methods, we will refer to both as Gradient Methods (GM) in the rest of this chapter.

4.4 Fast Gradient Method with (δ, L)-oracle

4.4.1 Convergence analysis

Let $\{\alpha_k\}_{k=0}^{\infty}$ be a sequence of reals such that

$$\alpha_0 \quad \in \quad (0, 1], \quad \alpha_k^2 \ \leq \ A_k \ \stackrel{\text{def}}{=} \ \sum_{i=0}^{k} \alpha_i, \quad k \geq 0. \tag{4.35}$$

and $\{\tau_k\}_{k \geq 0}$ be defined by $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.

Let us choose a setup and consider the FGM, developed in [59] (see section 2.4.5), but used here with a (δ, L) oracle:

Algorithm 15 Fast Gradient Method (FGM)

1: Choose $\alpha_0 \in (0,\ 1]$ and $x_0 = \min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$
4: Compute $y_k = T_L(x_k, g_{\delta,L}(x_k))$
5: Compute $z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle\}$
6: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
7: **end for**

Denote $\psi_k^* = \min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i \rangle]\}$.

Lemma 4.7. *For all $k \geq 0$, we have $A_k f(y_k) \leq \psi_k^* + E_k$ with $E_k = \sum_{i=0}^{k} A_i \delta$.*

Proof. Denote $f_k = f_{\delta,L}(x_k)$, and $g_k = g_{\delta,L}(x_k)$. For $k = 0$, we have

$$\begin{aligned} \psi_0^* &= \min_{x \in Q} \{Ld(x) + \alpha_0[f_0 + \langle g_0, x - x_0\rangle]\} \\ &\overset{(2.8)}{\geq} \alpha_0 \min_{x \in Q} \left\{ f_0 + \langle g_0, x - x_0\rangle + \tfrac{L}{2}\|x - x_0\|_E^2 \right\} \overset{(4.2)}{\geq} \alpha_0 f(y_0) - \delta. \end{aligned}$$

Assume now that the statement of the theorem is true for some $k \geq 0$. Optimality conditions for the optimization problem solved defining z_k imply

$$\langle L\nabla d(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i g_i, x - z_k\rangle \quad \geq \quad 0, \quad \forall x \in Q.$$

Hence, in view of strong convexity of d,

$$\begin{aligned} Ld(x) &\geq Ld(z_k) + \langle L\nabla d(z_k), x - z_k\rangle + \tfrac{L}{2}\|x - z_k\|_E^2 \\ &\geq Ld(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i \langle g_i, z_k - x\rangle + \tfrac{L}{2}\|x - z_k\|_E^2 . \end{aligned}$$

Thus, we have for all $x \in Q$

$$\begin{aligned} & Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i\rangle] \\ \geq\ & Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i\rangle] \\ & + \tfrac{L}{2}\|x - z_k\|_E^2 + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle]. \end{aligned}$$

We have obtained

$$\psi_{k+1}^* \quad \geq \quad \psi_k^* + \min_{x \in Q}\{ \tfrac{L}{2}\|x - z_k\|_E^2 + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle]\}.$$

On the other hand, using our recurrence assumption $A_k f(y_k) \leq \Psi_k^* + E_k$, we have

$$\begin{aligned} & \psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ & \geq\ A_k f(y_k) - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ & \overset{(4.2)}{\geq}\ A_k[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ & =\ A_{k+1} f_{k+1} + \langle g_{k+1}, A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle - E_k. \end{aligned}$$

Taking into account that

$$A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})$$

$$= A_k\tau_k(y_k - z_k) + \alpha_{k+1}x - \alpha_{k+1}\tau_k z_k - \alpha_{k+1}(1-\tau_k)y_k = \alpha_{k+1}(x - z_k),$$

we obtain

$$\psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \geq A_{k+1}f_{k+1} + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle - E_k.$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &\geq A_{k+1}f_{k+1} - E_k + \min_{x\in Q}\{\tfrac{L}{2}\|x - z_k\|_E^2 + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle\} \\
&= A_{k+1}\left[f_{k+1} + \min_{x\in Q}\{\tfrac{L}{2A_{k+1}}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}\right] - E_k \\
&\overset{(4.35)}{\geq} A_{k+1}\left[f_{k+1} + \min_{x\in Q}\{\tfrac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}\right] - E_k.
\end{aligned}$$

For $x \in Q$, define $y = \tau_k x + (1-\tau_k)y_k$. Since $y - x_{k+1} = \tau_k(x - z_k)$, we obtain

$$\begin{aligned}
&\min_{x\in Q}\left\{\tfrac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\right\} \\
&= \min_{y}\left\{\tfrac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle : \ y \in \tau_k Q + (1-\tau_k)y_k\right\} \\
&\geq \min_{y\in Q}\left\{\tfrac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle\right\}.
\end{aligned} \tag{4.36}$$

Therefore,we have:

$$\begin{aligned}
\Psi_{k+1}^* &\geq A_{k+1}\left[f_{k+1} + \min_{x\in Q}\{\tfrac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}\right] - E_k \\
&\overset{(4.2),(4.36)}{\geq} A_{k+1}f(y_{k+1}) - E_k - A_{k+1}\delta_{k+1},
\end{aligned}$$

and we get $A_{k+1}f(y_{k+1}) \leq \Psi_{k+1}^* + E_{k+1}$ with $E_{k+1} = E_k + A_{k+1}\delta_{k+1}$. □

Theorem 4.8. *For all $k \geq 0$, we have $f(y_k) - f^* \leq \frac{1}{A_k}\left(Ld(x^*) + \sum_{i=0}^k A_i\delta\right)$.*

Proof. Denote $f_i = f_{\delta,L}(x_i)$, and $g_i = g_{\delta,L}(x_i)$. Then

$$\begin{aligned}\psi_k^* &= \min_{x\in Q}\left\{Ld(x) + \sum_{i=0}^{k}\alpha_i[f_i + \langle g_i, x - x_i\rangle]\right\} \\ &\leq Ld(x^*) + \sum_{i=0}^{k}\alpha_i[f_i + \langle g_i, x^* - x_i\rangle] \leq Ld(x^*) + A_k f(x^*).\end{aligned}$$

The proof now simply follows from the recurrence established in Lemma 4.7. □

A simple choice for the sequence $\{\alpha_i\}$ consists in letting $\alpha_i = \frac{i+1}{2}$ for which we have $A_k = \frac{(k+1)(k+2)}{4}$, $\tau_k = \frac{2}{k+3}$, and therefore

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)^2} + \frac{1}{(k+1)(k+2)}\sum_{i=0}^{k}(i+1)(i+2)\delta.$$

Since $\sum_{i=0}^{k}(i+1)(i+2) = \frac{1}{6}(k+1)(k+2)(2k+6)$, we obtain

Theorem 4.9. *Assume that f is endowed with a (δ, L)-oracle. Then the sequence y_k generated by the FGM satisfies*

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)(k+2)} + \frac{1}{6}(2k+6)\delta \leq \frac{2LR^2}{(k+1)^2} + \frac{1}{3}(k+3)\delta \quad (4.37)$$

where $R \stackrel{\text{def}}{=} \sqrt{2d(x^*)}$.

Remark 4.6. The same result can be obtained for the variant of FGM using Bregman distance. See Theorem 7.8 in Chapter 3 with $\phi = f$, $\sigma = 0$, $\beta_i = L$ and $\alpha_i = \frac{i+1}{2}$ for all $i \geq 0$.

4.4.2 Error accumulation

Contrarily to the classical gradient methods, the use of inexact oracle in FGM results in error accumulation. Indeed, while the first term in (4.37) decreases as $O(\frac{1}{k^2})$, the second term is increasing in k, and this FGM used with an inexact oracle is asymptotically divergent. Section 4.8 will prove that error accumulation and divergence are unavoidable for all fast first-order methods.

We now study the non-asymptotic behavior of FGM, and consider two cases.

a. Oracle accuracy δ is fixed. In this case, we can find the number of iterations k^* that achieves the minimal guaranteed residual for the objective function. Denote accuracy achieved after k iterations with

$$e(k) = \frac{4Ld(x^*)}{(k+1)^2} + \frac{1}{3}(k+1)\delta + \frac{2}{3}\delta.$$

Function $e(k)$ is convex in k and its minimum is reached at iteration k^*

$$k^* = 2\sqrt[3]{\frac{3Ld(x^*)}{\delta}} - 1$$

for which the guaranteed accuracy for the objective function is

$$e(k^*) = \Theta(L^{1/3}R^{2/3}\delta^{2/3}).$$

b. Oracle accuracy δ can be chosen. Let us assume that parameter L of the inexact oracle is independent on δ. If we need to reach accuracy ϵ for the residual $f(y_k) - f^*$, it is enough to perform k iterations, with k satisfying two inequalities:

$$\frac{4Ld(x^*)}{(k+1)^2} \leq \frac{\epsilon}{2}, \quad \frac{1}{3}(k+3)\delta \leq \frac{\epsilon}{2}.$$

The first inequality gives us $k \geq \sqrt{\frac{8Ld(x^*)}{\epsilon}} - 1$, and the second one gives $k \leq \frac{3\epsilon}{2\delta} - 3$. Therefore attaining both $\frac{\epsilon}{2}$ accuracies is possible if and only if

$$\delta \leq \frac{3\epsilon^{3/2}}{2\sqrt{8Ld(x^*)}+4\sqrt{\epsilon}}. \tag{4.38}$$

In conclusion, if we choose the oracle accuracy satisfying relation (4.38), then after

$$k(\epsilon) = \sqrt{\frac{8Ld(x^*)}{\epsilon}} - 1$$

iterations, we obtain a point $y_{k(\epsilon)} \in Q$ satisfying $f(y_{k(\epsilon)}) - f^* \leq \epsilon$.

We observe that, compared to GM, FGM requires a higher-order accuracy for the oracle ($\mathcal{O}(\epsilon^{3/2})$ versus $\mathcal{O}(\epsilon)$ for GM).

4.5 Comparison between classical and fast gradient methods

When an exact oracle is used, FGM is an optimal method for the class $F_L^{1,1}(Q)$. It reaches an objective function accuracy ϵ after $O\left(\sqrt{\frac{L}{\epsilon}}R\right)$ iterations while GM requires $O\left(\frac{LR^2}{\epsilon}\right)$ iterations for the same result.

Performing such a comparison becomes more complicated when an inexact first-order oracle is used. Contrary to GM, FGM suffers from error accumulation. In order to compare their efficiency, we consider two cases.

4.5.1 Oracle accuracy δ can be freely chosen

In this case we assume that L is independent from the oracle accuracy δ (see examples in Section 4.2). If we need to reach ϵ accuracy for the objective function, GM will work using an inexact oracle with $\delta = \Theta(\epsilon)$. However, it will then need $O\left(\frac{LR^2}{\epsilon}\right)$ iterations.

For FGM with inexact oracle, error accumulation forces the use of a more accurate oracle, i.e. with $\delta = \Theta\left(\frac{\epsilon^{3/2}}{\sqrt{L}R}\right)$. However only $O\left(\sqrt{\frac{L}{\epsilon}}R\right)$ iterations are needed. Thus, the choice between two methods depends on the complexity of the inexact oracle. Denote by $C(\delta)$ the cost associated with computing an answer $(f_{\delta,L}(x), g_{\delta,L}(x))$ for a (δ, L) inexact oracle. We see that GM is preferable to FGM if the following holds (up to constant factors in the arguments of $C(\cdot)$)

$$\frac{1}{\epsilon}LR^2C(\epsilon) \quad < \quad \frac{1}{\epsilon^{1/2}}L^{1/2}RC\left(\frac{\epsilon^{3/2}}{L^{1/2}R}\right)$$

which leads us to consider the following situations:

- ◇ Oracle for which higher accuracy is very expensive: $C(\delta) = \Omega\left(\frac{1}{\delta}\right)$ (e.g. $C(\delta) = \frac{1}{\delta^2}$). In this case, it is preferable to use GM.
- ◇ Oracle for which higher accuracy is moderately expensive: $C(\delta) = \Theta\left(\frac{1}{\delta}\right)$. For such an oracle both methods are equivalent.
- ◇ Oracle for which higher accuracy is cheap: $C(\delta) = o\left(\frac{1}{\delta}\right)$ (for example, $C(\delta) = \frac{1}{\delta^{1/2}}$, or even $C(\delta) = \ln\frac{1}{\delta}$). FGM is here better than GM.

4.5.2 Oracle accuracy δ is fixed.

In this case, the sequence of iterates generated by GM satisfies inequality

$$f(y_k) - f^* \quad \leq \quad \frac{LR^2}{2k} + \delta,$$

whereas the sequence obtained by FGM satisfies inequality

$$f(y_k) - f^* \quad \leq \quad \frac{2LR^2}{(k+1)(k+2)} + \frac{k+3}{3}\delta.$$

Figures 2, 3 and 4 depict these two rates of convergence for three different values of the oracle accuracy parameter δ (with $L = R = 1$ in all cases).

The higher the accuracy of the oracle, the larger the threshold in number of iterations after which FGM is better than GM. For example, on Figure 4, we

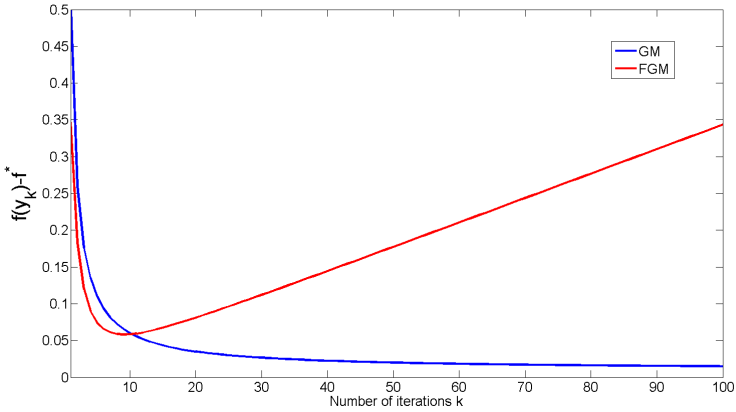


Figure 2: Convergence rate of GM and FGM with $\delta = 1e-2$, $L = 1$ and $R = 1$

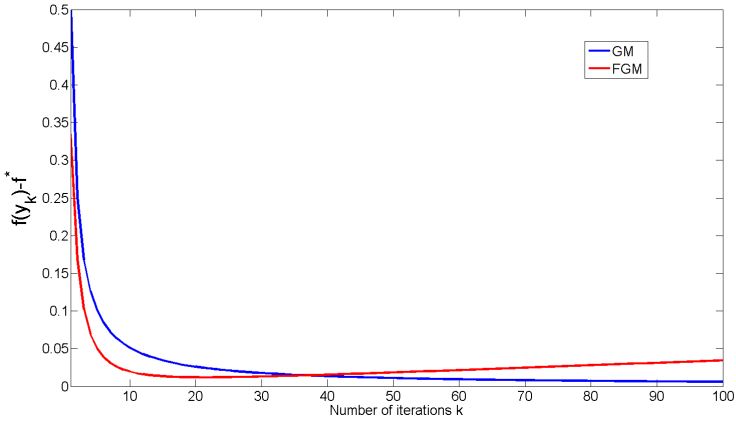


Figure 3: Convergence rate of GM and FGM with $\delta = 1e-3$, $L = 1$ and $R = 1$

see that when the oracle accuracy is sufficiently high ($\delta = 1e-4$), FGM outperforms GM accuracy at least for the first hundred iterations. In the exact case, i.e. for oracle accuracy $\delta = 0$, FGM outperforms GM for any number of iterations.

On the other hand, when oracle accuracy is low, accumulation of oracle errors in FGM becomes so prevalent that GM is better than FGM, except for the first

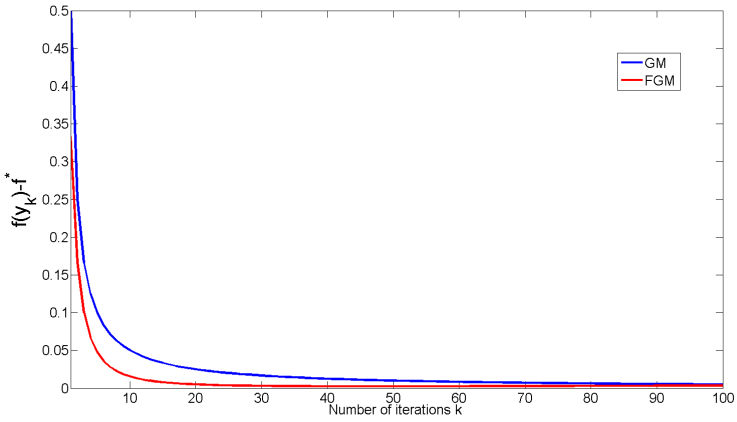


Figure 4: Convergence rate of GM and FGM with $\delta = 1e-4$, $L = 1$ and $R = 1$

few iterations. Figure 2 (with $\delta = 1e-2$) depicts this situation.

For intermediate values of accuracy (such as on Figure 3, where $\delta = 1e-3$), the situation is more complicated. During the first iterations, FGM reduces errors much better than GM, because of its better convergence rate. For FGM, the error attains its minimum value after $N_1 = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ iterations, with corresponding accuracy $\epsilon^*_{FGM} = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$. It is not interesting to perform further FGM iterations since the gap can then only increase due to error accumulation.

Note that there exists an iteration threshold N_2 ($> N_1$) after which GM provides better accuracy than FGM. However, this does not mean that GM is superior to FGM as soon as we reach that number of iterations, because FGM already achieved a lower accuracy ϵ^*_{FGM} after N_1 iterations. If we wait further until we reach $N_3 = \Theta\left(\frac{LR^2}{\delta^{2/3}}\right)$ ($> N_2$) iterations, the accuracy of GM finally becomes better than ϵ^*_{FGM}, the best reachable accuracy with FGM. Final accuracies ϵ between ϵ^*_{FGM} and δ can then only be reached by GM (they are inaccessible by FGM), and require $\Theta\left(\frac{LR^2}{\epsilon-\delta}\right)$ iterations.

In conclusion, FGM is the method of choice when we need accuracy not better than $\epsilon^*_{FGM} = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$. Indeed, accuracy ϵ^*_{FGM} is reached by the FGM after N_1 iterations whereas the GM needs N_3 iterations in order to obtain

the same error. In order to obtain accuracy better than ϵ^*_{FGM}, GM must be used since the FGM cannot decrease the error below ϵ^*_{FGM}.

4.6 Comparison with other types of inexact oracle

Fast-gradient methods using inexact first-order oracle have been recently studied in [21] and [2]. These works assume that set Q is bounded and that the oracle provides at each point $y \in Q$ an approximate gradient $g(y)$ satisfying condition

$$|\langle g(y) - \nabla f(y), x - z\rangle| \quad \leq \quad \xi \quad \forall x, y, z \in Q. \tag{4.39}$$

Let us compare this definition with (4.2), taking into account both their applicability and the results obtained. First of all, the existence of an inexact oracle satisfying (4.39) require more assumptions than our definition:

- ◇ Set Q must be bounded (this is not needed for (4.2)).
- ◇ Objective function f must be differentiable. The existence of the gradient at all points is necessary since it must be compared with the approximate gradient. Our approach is also able to consider non- or weakly smooth convex functions.

Furthermore, even in the smooth case $f \in F_L^{1,1}(Q)$ with bounded Q, we argue that condition (4.39) is strictly stronger than (4.2). Assume $f \in F_L^{1,1}(Q)$.

1. Any approximate gradient $g(y)$ satisfying (4.39) also satisfies our definition. Indeed, in view of (4.1) and (4.39), we have for all $x, y \in Q$

   $$f(y) - \xi + \langle g(y), x - y\rangle \leq f(x) \leq f(y) + \xi + \langle g(y), x - y\rangle + \tfrac{L}{2}\|x - y\|_E^2 .$$

   and therefore taking $f_{\delta,L}(y) = f(y) - \xi$, and $g_{\delta,L}(y) = g(y)$ satisfies (4.2) with $\delta = 2\xi$ and the same value for L.

2. On the other hand, our condition (4.2) does not imply (4.39) with any $\xi = \Theta(\delta)$. Indeed, consider the function $f(x) = \max_{u\in U} \Psi(x, u)$, where

   $$\Psi(x, u) = -\tfrac{1}{2}\|u\|_2^2 + \langle x, u\rangle, \quad Q = \{y \in \mathbb{R}^n : \|y\|_2 \leq 1\}, \quad U = \mathbb{R}^n. \tag{4.40}$$

   Let us assume the answer of oracle for $x = 0$ is obtained for some point u_0 satisfying $\|u_0\|_2 = \delta^{1/2}$. Since $u_0^* = \arg\max_{u\in U} \Psi(0, u) = 0$, and $f(0) - \Psi(0, u_0) = \frac{1}{2}\|u_0\|_2^2 = \frac{\delta}{2}$, the pair $(f_{\delta,L}(0), g_{\delta,L}(0)) = (-\frac{\delta}{2}, u_0)$ is

an acceptable answer for a (δ, L) inexact oracle with $L = 2$ (see subsection 4.2.2). However we can check that

$$\max_{y,z \in Q} |\langle \nabla f(0) - g_{\delta,L}(0), y - z \rangle| = 2 \max_{y \in Q} |\langle u_0, y \rangle| = 2\delta^{1/2}.$$

We now compare efficiency estimates of FGM based on these oracles. FGM using oracle (4.39) converges as follows:

$$f(y_k) - f^* \quad \leq \quad \frac{CLR^2}{k^2} + 3\xi,$$

where C is an absolute constant. This bound does not feature error accumulation, meaning the accuracy ξ of the oracle can be chosen to be of the same order as the desired accuracy ϵ of the solution. This result seems at first sight to be better than what we obtained with our (δ, L)-oracle.

However, we noted that for the same level of accuracy, condition (4.39) is much stronger than (4.2). Let us look at important example. Consider the class of functions with explicit max-structure: $f(x) = \max_{u \in U} \Psi(x, u)$, where set U is closed and convex, and $\Psi(x, u) = G(u) + \langle x, Au \rangle$, where $G(u)$ is a differentiable, strongly concave function with concavity parameter κ. Assume that we want to solve the primal problem $\min_{x \in Q} f(x)$ with accuracy ϵ. With our definition of inexact oracle, the oracle accuracy δ corresponds directly to the (objective function) accuracy required when solving the dual problem (see subsection 4.2.2).

In the case of an approximate gradient satisfying definition (4.39), we can also use an approximate dual solution $u_x \approx u_x^*$

$$\nabla f(x) \quad = \quad Au_x^*, \quad g(x) = Au_x.$$

However, we need to satisfy the following relation:

$$|\langle A(u_x^* - u_x), y - z \rangle| \quad \leq \quad \epsilon, \quad \forall x, y, z \in Q. \tag{4.41}$$

(We can take $\xi = \epsilon$ since the condition (4.39) avoids accumulation of errors). For that, we need to have u_x close to u_x^* according to

$$\|u_x - u_x^*\|_F \quad \leq \frac{\epsilon}{\operatorname{diam}(Q) \cdot \|A\|_{F \to E^*}}.$$

Since Ψ is strongly concave, i.e. $\Psi(x, u_x^*) - \Psi(x, u_x) \geq \frac{\kappa}{2} \|u_x - u_x^*\|_F^2$, a sufficient condition for (4.39) is then as follows

$$\Psi(x, u_x^*) - \Psi(x, u_x) \leq \frac{\kappa}{2} \left(\frac{\epsilon}{\operatorname{diam}(Q) \cdot \|A\|_{F \to E^*}} \right)^2 = \mathcal{O}(\epsilon^2).$$

Compare this to our approach, for which it was enough to solve the dual problem up to accuracy $\epsilon^{3/2}$ (see (4.38)) in order to avoid accumulation of errors.

Remark 4.7. In some cases, inequality $\Psi(x, u_x^*) - \Psi(x, u_x) \leq \epsilon^2/8$ is also a necessary condition for (4.41). Indeed, consider again the saddle point problem defined by (4.40). We have $f(0) - \Psi(0, u_0) = \frac{1}{2}\|u_0\|_2^2$. In order to satisfy condition (4.41) we need to ensure

$$\epsilon \;\geq\; 2\max_{y \in Q} |\langle u_0, y\rangle| \;=\; 2\|u_0\|_2 = 2\sqrt{2(f(0) - \Psi(0, u_0))}.$$

Remark 4.8. The definition of inexact oracle used in [2] is slightly different from (4.39). The author considers functions that are not necessarily differentiable, and assumes that his first-order oracle $g(y)$ satisfies the following conditions:

$$\begin{aligned} f(x) &\geq f(y) + \langle g(y), x - y\rangle - \bar{\xi} \quad \forall x \in \operatorname{dom} f \\ f(x) &\geq f(y) + \langle g(y), x - y\rangle - \bar{\xi}\,\|x - y\| \quad \forall x \in \operatorname{dom} f \end{aligned}$$

and that the set Q is bounded.

He also shows (see Lemma 5.4 in [2]) that the above second condition implies $\|g(y) - g_y\|^* \leq \bar{\xi}$, where $g_y \in \partial f(y)$. It is easy to see that this implies (4.39) with $\xi = D_Q\bar{\xi}$ (where D_Q denotes the diameter of Q), possibly replacing $\nabla f(y)$ with a subgradient when function f is nonsmooth.

4.7 First-order methods of smooth convex optimization applied to functions with lack of smoothness

4.7.1 Solving weakly smooth problems

Let f be a convex function satisfying the Hölder condition (4.12). This class includes nonsmooth convex functions with bounded variation of subgradients ($\nu = 0$), and smooth convex functions with Hölder continuous gradient ($\nu \in (0, 1]$). We have shown in subsection 4.1.3, that for all $\delta > 0$ these functions can be equipped with a (δ, L)-oracle with

$$L \;=\; A(\delta, \nu) \;=\; L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}.$$

This observation allows us to apply first-order methods of $F_L^{1,1}(Q)$ to functions with weaker level of smoothness, replacing the gradients by subgradients and using a Lipschitz constant L that grows as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$ in terms of the δ parameter of the oracle.

This parameter δ, which does not correspond to the actual accuracy of the oracle, will have to be properly tuned in numerical methods, with a trade-off between the "accuracy" of the oracle, and the Lipschitz constant L.

For the sake of simplicity, we assume in the rest of this section that a fixed number of iterations N is performed.

Let us apply GM to a weakly smooth function f with an inexact (δ, L)-oracle. In view of (4.31), after N iterations we have

$$f(y_N) - f^* \quad \leq \quad L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{2N} + \delta \quad \stackrel{\text{def}}{=} \quad C_N \left(\tfrac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} + \delta.$$

Denote $\tau = \frac{1-\nu}{1+\nu}$. Then the optimal accuracy δ_N can be found from the equation

$$C_N \tfrac{\tau}{\delta_N^{1+\tau}} \quad = \quad 1.$$

Thus, we come to the following bound:

$$f(y_N) - f^* \quad \leq \quad \delta_N \left(\tfrac{C_N}{\delta_N^{1+\tau}} + 1\right) \; = \; \tfrac{2\delta_N}{1-\nu}. \tag{4.42}$$

Note that

$$\begin{aligned} \delta_N \quad &= \quad (\tau C_N)^{\frac{1}{1+\tau}} \\ &= \quad \left(\tfrac{1-\nu}{1+\nu} \cdot L_\nu \left[\tfrac{L_\nu}{2} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{2N}\right)^{\frac{1+\nu}{2}} \\ &= \quad \tfrac{1-\nu}{1+\nu} \cdot \tfrac{L_\nu R^{1+\nu}}{2^{\frac{1-\nu}{2}} \cdot N^{\frac{1+\nu}{2}}} \end{aligned}$$

and

$$L = \frac{L_\nu N^{\frac{1-\nu}{2}}}{R^{1-\nu} 2^{\frac{1-\nu}{2}}}.$$

Thus, we come to the following upper bound:

$$f(y_N) - f^* \quad \leq \quad \tfrac{L_\nu R^{1+\nu}}{1+\nu} \cdot \left(\tfrac{2}{N}\right)^{\frac{1+\nu}{2}}. \tag{4.43}$$

For functions with bounded variation of subgradients ($\nu = 0$), we get:

$$f(y_N) - f^* \quad \leq \quad L_0 R \cdot \left(\tfrac{2}{N}\right)^{\frac{1}{2}},$$

which is the optimal rate of convergence (see [55, 58]). However for functions with Hölder continuous gradient ($0 < \nu$), the obtained rate is not optimal (it should be $O(N^{-\frac{1+3\nu}{2}})$, see [53, 39]).

Let us now apply FGM to a weakly smooth function using an (δ, L)-oracle. In view of (4.37), after N iterations we have:

$$\begin{aligned} f(y_N) - f^* &\leq 4L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{d(x^*)}{(N+1)^2} + \delta \cdot (\tfrac{N}{3} + 1) \\ &\leq 2L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{(N+1)^2} + \delta \cdot (N+1) \\ &\stackrel{\text{def}}{=} \hat{C}_N \left(\tfrac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} + \delta \cdot (N+1). \end{aligned}$$

The equation for optimal δ_N now becomes $\hat{C}_N \frac{\tau}{\delta_N^{1+\tau}} = N+1$. Therefore, we get

$$f(y_N) - f^* \leq \delta_N \left(\tfrac{\hat{C}_N}{\delta_N^{1+\tau}} + N + 1\right) = \tfrac{2\delta_N}{1-\nu}(N+1).$$

Note that

$$\begin{aligned} \delta_N &= (\hat{C}_N \tfrac{\tau}{N+1})^{\frac{1}{1+\tau}} = \left(\tfrac{1-\nu}{1+\nu} \cdot 2L_\nu \left[\tfrac{L_\nu}{2} \cdot \tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{(N+1)^3}\right)^{\frac{1+\nu}{2}} \\ &= \tfrac{1-\nu}{1+\nu} \cdot \tfrac{L_\nu R^{1+\nu}}{(N+1)^{\frac{3}{2}(1+\nu)}} \cdot 2^\nu \end{aligned}$$

and

$$L = \frac{L_\nu (N+1)^{\frac{3(1-\nu)}{2}}}{2^{1-\nu} R^{1-\nu}}.$$

Thus, we obtain the following upper bound

$$f(y_N) - f^* \leq \tfrac{2^{1+\nu} L_\nu R^{1+\nu}}{1+\nu} \left(\tfrac{1}{N+1}\right)^{\frac{1+3\nu}{2}}. \tag{4.44}$$

For functions with bounded variation of subgradients ($\nu = 0$), we get

$$f(y_N) - f(x^*) \leq 2L_0 R \left(\tfrac{1}{N+1}\right)^{\frac{1}{2}}.$$

In all cases, we obtain the optimal rate of convergence. Therefore, FGM can be seen as a universal first-order method simultaneously optimal for smooth, weakly smooth and nonsmooth convex functions provided that the steplength is chosen according to $\frac{1}{L} = \frac{2^{1-\nu} R^{1-\nu}}{L_\nu (N+1)^{\frac{3(1-\nu)}{2}}}$.

The applicability of first-order method of smooth convex optimization to nonsmooth convex problems, justified by the notion of (δ, L)-oracle, has several further interesting consequences. We describe two of them below.

- ◇ We can apply GM and FGM to objective functions composed of a sum of smooth and nonsmooth components.
- ◇ We can get lower bounds on the rate of accumulation of errors in the first-order methods based on (δ, L)-oracle. It appears that error accumulation is an intrinsic property of any fast gradient method.

4.7.2 Solving composite optimization problems

Consider the composite convex objective function:

$$f(x) = f_1(x) + f_2(x),$$

where f_1 is a smooth convex function with Lipschitz continuous gradient (constant $L(f_1)$), and f_2 is a nonsmooth convex function with subgradients whose variation is bounded by constant $M(f_2)$. We assume that the standard exact first-order oracles are available for both f_1 and f_2.

Note that function f_1 is equipped with $(0, L(f_1))$-oracle, and by (4.15) function f_2 has $(\delta, \frac{1}{2\delta}M^2(f_2))$-oracle. Hence, we conclude that the pair

$$(f_1(y) + f_2(y), \nabla f_1(y) + g_2(y)), \quad g_2(y) \in \partial f_2(y), \tag{4.45}$$

is a (δ, L)-oracle for function f with $L = L(f_1) + \frac{1}{2\delta}M^2(f_2)$. Assume again that the number of iterations N for our methods is fixed.

Let us apply now GM to function f using the inexact (δ, L)-oracle (4.45). Then, after N iterations we have:

$$f(y_N) - f^* \overset{(4.31)}{\leq} \left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right)\tfrac{R^2}{2N} + \delta.$$

Minimizing this expression with respect to $\delta \geq 0$, we obtain $\delta^* = \frac{M(f_2)R}{2N^{1/2}}$. Therefore, the best upper bound for the residual is

$$f(y_N) - f^* \leq \tfrac{L(f_1)R^2}{2N} + \tfrac{M(f_2)R}{N^{1/2}}.$$

This method has the optimal rate of convergence for nonsmooth part of the problem, but not for the smooth one.

Let us check now the performance of FGM as applied to the composite problems. In view of (4.37), we have after N iterations of the scheme

$$\begin{aligned} f(y_N) - f^* &\leq 4\left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right)\tfrac{d(x^*)}{(N+1)^2} + \delta \cdot (\tfrac{N}{3} + 1) \\ &\leq 2\left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right)\tfrac{R^2}{(N+1)^2} + \delta \cdot (N+1). \end{aligned}$$

Minimizing this function in $\delta \geq 0$, we obtain: $\delta^* = \frac{M(f_2)R}{(N+1)^{3/2}}$. The upper-bound therefore becomes

$$f(y_N) - f^* \leq \tfrac{2L(f_1)R^2}{(N+1)^2} + \tfrac{2M(f_2)R}{(N+1)^{1/2}}.$$

For such a composite objective function, this method is optimal both for the smooth and nonsmooth parts of the problem.

Remark 4.9. Our analysis is in a certain sense similar to that of [44], where the author applies a version of FGM to a stochastic composite optimization problem. In the deterministic case, the author applies a variant of FGM, replacing the gradients by subgradients in the nonsmooth part of objective, and the Lipschitz constant by a quantity of order $\mathcal{O}(M(f_2)N^{3/2})$. This method appears to be optimal both for the smooth and nonsmooth parts of the composite function. In our approach, $N = \Theta((\frac{1}{\delta}M(f_2))^{2/3})$, and we get $M(f_2)N^{3/2} = \Theta(\frac{1}{\delta}M^2(f_2))$, which is, up to a constant factor, the quantity that replaces the Lispchitz constant for our method.

4.8 First-order methods and error accumulation

Applicability of first-order methods of smooth optimization to nonsmooth problems, based on the notion of inexact oracle, opens a possibility to derive lower bounds on error accumulation. This is the main subject of this section. We recall first the lower bound for the complexity of any first-order method for nonsmooth convex problems:

Theorem 4.10. *(see [55]) Let $\mathcal{M}$ be a first-order method able to solve any nonsmooth convex problem of the form $\min_{x\in Q} f(x)$ with $f \in F_M^{0,0}(Q)$. Then the complexity of the method $\mathcal{M}$ cannot be better than $\frac{M^2R^2}{\epsilon^2} - 1$ (where $R = \|x_0 - x^*\|$) i.e. the general convergence rate of a first-order method for nonsmooth convex problems cannot be better than $\frac{MR}{\sqrt{k+1}}$.*

Using this lower-bound, we are able to obtain the following result, linking rate of convergence and rate of error accumulation:

Theorem 4.11. *Consider a first-order method with convergence rate $LR^2\Xi_1(k)$ when exact first-order information is used. Assume that the bounds on the performance of this method, applied to a problem equipped with an inexact (δ, L)-oracle, are given by inequality*

$$f(y_k) - f^* \leq LR^2\Xi_1(k) + \delta\Xi_2(k) \tag{4.46}$$

where k is the iteration counter. Then the inequality

$$\Xi_1(k)\Xi_2(k) \geq \frac{1}{2(k+1)}$$

must hold for all $k \geq 0$.

Proof. Let f be a nonsmooth convex function, whose subgradients have variation bounded by constant M. We have seen that for such a function, the

standard oracle can be treated as $(\delta, \frac{M^2}{2\delta})$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$f(y_k) - f^* \leq \frac{M^2 R^2}{2\delta}\Xi_1(k) + \delta\Xi_2(k).$$

Optimizing the right-hand side of this inequality in δ, we get

$$f(y_k) - f^* \leq \sqrt{2}MR\sqrt{\Xi_1(k)\Xi_2(k)}.$$

From the lower complexity bounds for nonsmooth optimization problems (see Theorem 4.10), we know that black-box methods cannot converge faster than $\frac{MR}{\sqrt{k+1}}$. Hence, we conclude that $\Xi_1(k)\Xi_2(k) \geq \frac{1}{2(k+1)}$.

□

This result show us that the rate of convergence $\Xi_1(k)$ and the rate of error accumulation $\Xi_2(k)$ are linked: $\Xi_2(k)$ must grow with a rate at least of order $\Theta\left(\frac{1}{k\Xi_1(k)}\right)$.

Considering the particular case where $\Xi_1(k) = \frac{C_1}{k^p}$ and $\Xi_2(k) = C_2 k^q$, we obtain:

Theorem 4.12. *Consider a first-order method for $F_L^{1,1}(Q)$ with convergence rate $O(\frac{LR^2}{k^p})$ when exact first-order information is used. Assume that the bounds on the performance of this method applied to a problem equipped with an inexact (δ, L)-oracle are given by inequality*

$$f(y_k) - f^* \leq \frac{C_1 L R^2}{k^p} + C_2 k^q \delta, \tag{4.47}$$

where C_1, C_2 are absolute constants. Then the inequality $q \geq p-1$ must hold.

In the exact case, when minimizing a function in $F_L^{1,1}(Q)$, any first-order method with convergence rate $\Theta(\frac{LR^2}{k^2})$ is optimal (e.g. FGM), and any method with convergence rate $\Theta(\frac{LR^2}{k})$ is suboptimal (e.g. GM). In the case of inexact (δ, L)-oracle, the situation is more complicated.

Total performance of the method also depends from the way it accumulates successive errors coming from the oracle. In this situation, the superiority of FGM over GM is not completely clear anymore. As we have seen in the previous sections, FGM suffers from accumulation of errors, but GM does not.

From Theorem 4.12, we know that this accumulation is a direct consequence of the fast convergence of the scheme. Any method with complexity estimate $\Theta(\sqrt{\frac{L}{\epsilon}}R)$ must suffer from this instability. It appears that in the inexact situation, both FGM and GM are optimal, but in different senses.

- $q = 0 \Rightarrow p \leq 1$:
  It is impossible to have a first-order method without accumulation of errors, which has better complexity than GM, that is $\Theta(\frac{LR^2}{\epsilon})$.
- $p = 2 \Rightarrow q \geq 1$:
  On the other hand, if we have a first-order method with complexity $\Theta(\sqrt{\frac{L}{\epsilon}}R)$, then it must be accumulating errors, which grow at least as $\Theta(k\delta)$.

The next theorem relates the rate of convergence of the method with the required accuracy of the oracle.

Theorem 4.13. *Let parameter L of inexact oracle (4.2) be independent from δ. Under assumptions of Theorem 4.12, accuracy ϵ in the residual of the objective function requires at least the following accuracy of the oracle:*

$$\delta \quad \leq \quad \frac{p\cdot\epsilon}{(p+q)C_2}\left[\frac{q\cdot\epsilon}{(p+q)C_1LR^2}\right]^{q/p}.$$

Proof. In order to guarantee accuracy ϵ by the estimate (4.47), we have to choose k and δ such that:

$$\frac{C_1LR^2}{k^p} \leq \alpha\epsilon, \quad C_2k^q\delta \leq (1-\alpha)\epsilon$$

for some $\alpha \in [0,1]$. The first inequality gives us $k \geq \left[\frac{C_1LR^2}{\alpha\epsilon}\right]^{1/p}$, and using the second inequality, we obtain

$$C_2\left[\frac{C_1LR^2}{\alpha\epsilon}\right]^{q/p}\delta \quad \leq \quad (1-\alpha)\epsilon.$$

Thus, $\delta \leq \frac{(1-\alpha)\alpha^{q/p}\cdot\epsilon^{(p+q)/p}}{C_2[C_1LR^2]^{q/p}}$. It remains to maximize the right-hand side of this inequality in α. □

Corollary 4.14. *If a first-order method has efficiency estimate $\Theta\left(\frac{LR^2}{\epsilon}\right)$, then it can be applied to a (δ, L)-oracle, with accuracy at least $\Omega(\frac{\epsilon^{1+q}}{L^qR^{2q}})$ or higher. For methods optimal with respect to accumulation of errors ($q = p-1 = 0$), like the GM, we can choose $\delta = \Omega(\epsilon)$.*

Corollary 4.15. *If a first-order method has efficiency estimate $\Theta\left(\sqrt{\frac{L}{\epsilon}}R\right)$, then it can be applied to a (δ, L)-oracle, with accuracy at least $\Omega(\frac{\epsilon^{1+q/2}}{L^{q/2}R^q})$ or higher. For methods optimal with respect to accumulation of errors ($q = p-1 = 1$), like the FGM, we can choose $\delta = \Omega(\frac{\epsilon^{3/2}}{L^{1/2}R})$.*

Remark 4.10. These different theorems linking the fastness and the sensitivity to errors of any first-order method depend of course on our definition of inexact information, the (δ, L)-oracle. Different results (more optimistic or pessimistic) can be obtained using other definitions of inexact oracle. Our choice of the (δ, L)-oracle is motivated at the same time by the various natural examples of inexact first-order information that fit with this definition (see subsection 4.1.3 and section 4.2) and by the fact that its effect on first-order methods does not depend on the boundedness of the feasible set.

Remark 4.11. Using the notion of approximate gradient
i.e. $\left\|\nabla f(x) - \tilde{\nabla} f_x\right\|_E^* \leq \Delta$ (see subsection 4.1.3) instead of the (δ, L)-oracle, the gradient method is convergent in the bounded case but divergent in the unbounded case in the absence of reliable stopping criterion.

Indeed, when the feasible set is bounded, an approximate gradient is a particular case of (δ, L) oracle and the gradient method is therefore convergent. When the feasible set is unbounded, it is possible to show that the gradient method can diverge. An easy example is given by $f(x) = \frac{L}{2}(x_{(1)})^2 + \frac{\Delta}{2}x_{(2)}$, $Q = \mathbb{R} \times [0, \infty[$ and $\tilde{\nabla} f_x = (Lx_{(1)}, -\frac{\Delta}{2})$.

These very different behaviors between the bounded and unbounded case is an important drawback of the notion of approximate gradient. The definition of approximate gradient, contrarily to the (δ, L)-oracle does not take into account the geometry of the feasible set and cannot therefore guarantee a specific behavior for a method independently of the feasible set boundedness.

Chapter 5

First-Order Methods with Inexact Oracle: the smooth strongly convex case

This chapter corresponds to the paper [26]:
O. Devolder, F. Glineur and Yu. Nesterov. **First-order methods with inexact oracle: the strongly convex case.** *CORE Discussion Paper 2013/16.*

Chapter 5 in two Questions/Answers.

◇ *Can we exploit the strong convexity of the objective function in order to reduce the sensitivity of the first-order methods with respect to oracle errors ?*

Even in the presence of strong convexity, fastness and sensitivity to errors are linked. As in the smooth convex case, the simple gradient methods (PGM/DGM) can be seen as robust but slower, whereas the FGM is faster but more sensitive to oracle errors. However, the strong convexity leads to much faster convergence rates (linear instead of sublinear) and to a reduced sensitivity with respect to oracle errors (bounded instead of unbounded accumulation of errors for the FGM). The central quantity is now the condition number $\frac{L}{\mu}$ of the smooth strongly convex objective function.

◇ *Is it possible to apply first-order methods, initially designed for smooth strongly convex problems, to strongly convex function with weaker level of smoothness or with different level of convexity? What do the complexities of the PGM/DGM and FGM become on these different classes*

of problems?

It appears that the notion of (δ, L, μ)-oracle, introduced in order to model lack of accuracy in the first-order information of a smooth strongly convex function, can be also used in order to represent a lack of smoothness or a lack of strong convexity. Using this trick, the first-order methods initially designed for smooth strongly convex functions can be applied to strongly convex functions with a weaker level of smoothness (nonsmooth strongly convex functions and weakly smooth strongly convex functions) but also to uniformly convex functions with various levels of smoothness (smooth, nonsmooth or weakly smooth). In view of this result, we are able to derive the corresponding complexities of the PGM/DGM and FGM on these different classes (see section 5.6).

Contents

5.1 The (δ, L, μ)-oracle **146**
5.1.1 Motivation and definition 146
5.1.2 Properties 146
5.2 Examples of (δ, L, μ)-oracle **148**
5.2.1 Strongly convex function with computation at shifted points 148
5.2.2 Functions approximated by a smooth strongly convex function 149
5.2.3 Strongly convex function with approximate function value and approximate gradient 149
5.2.4 Saddle-point functions 150
5.2.5 Uniformly convex functions with weaker level of smoothness 154
5.3 Primal Gradient Method with (δ, L, μ)-oracle **156**
5.4 Dual Gradient Method with (δ, L, μ)-oracle **158**
5.5 Fast Gradient Method with (δ, L, μ)-oracle **162**
5.5.1 The method 162
5.5.2 Convergence rate 163
5.5.3 Oracle accuracy fixed: Best reachable target accuracy using the FGM 172
5.5.4 Oracle accuracy not fixed: Required number of iterations and required oracle accuracy for a given target accuracy 174
5.6 Application to weakly smooth uniformly convex functions **178**
5.6.1 Gradient Method for function in $U_{\kappa,M}^{0,\rho,\nu}(Q)$. . . 178
5.6.2 Fast Gradient Method for functions in $U_{\kappa,M}^{0,\rho,\nu}(Q)$ 180
5.7 Lower bound on error increase **182**

In the previous chapter, we have studied the effect of inexact first-order information on the first-order methods of smooth convex optimization. In this chapter, we do the same but for the first-order methods designed for smooth strongly convex problems.

In Section 5.1, we introduce the notion of (δ, L, μ)-oracle, that can be seen as a extension of the (δ, L)-oracle for the strongly convex case, and establish some basic properties of such kind of oracles. In Section 5.2, we consider different examples of (δ, L, μ) oracle: strongly convex functions with first-order information computed at shifted points, strongly convex functions with approximate gradient, strongly convex max-functions with inexact resolution of subproblems, etc. We prove also that the notion of (δ, L, μ)-oracle can be used in order to model exact first-order information of weakly smooth uniformly convex functions.

Sections 5.3, 5.4 and 5.5 are devoted to the behavior analysis of three first-order methods of smooth strongly convex optimization, respectively the PGM, the DGM and the FGM, when used with a (δ, L, μ)-oracle. As in the smooth convex case, we obtain that the PGM (and the DGM) can be seen as robust but relatively slow methods, whereas the FGM is faster but more sensitive to oracle errors. However, strong convexity leads to much faster convergence rates (linear instead of sublinear) for every method and to a smaller sensitivity with respect to oracle errors (bounded instead of unbounded accumulation of errors for the FGM).

In Section 5.6, using the fact that an exact oracle of a weakly smooth uniformly convex function can be seen as a (δ, L, μ)-oracle, we obtain the complexity of our different first-order methods on such kind of objective function. The last section (Section 5.7) is devoted to the obtainment of lower bounds on the error increase for any first-order method designed for smooth strongly convex functions and used with a (δ, L, μ)-oracle.

Remark 5.1. In this chapter, we restrict ourselves to the Euclidean setup, assuming that the finite-dimensional vector space E is endowed with an Euclidean norm, defined for a given arbitrary positive definite self-adjoint operator $B : E \to E^*$ by

$$\|h\|_E = \|h\|_2 = \langle Bh, h\rangle^{1/2} \quad \forall h \in E.$$

5.1 The (δ, L, μ)-oracle

5.1.1 Motivation and definition

Consider $S^{1,1}_{\mu,L}(Q)$, the class of strongly convex functions (with parameter μ) on convex set Q whose gradient is Lipschitz-continuous (with constant L). It is well-known that functions belonging to this class satisfy

$$\frac{\mu}{2}\|x-y\|_E^2 \leq f(x) - (f(y) + \langle \nabla f(y), x-y\rangle) \leq \frac{L}{2}\|x-y\|_E^2, \quad \forall x, y \in Q. \tag{5.1}$$

Moreover, it is easy to check that, for a given y, quantities $f(y)$ and $\nabla f(y)$ are uniquely determined by this pair of inequalities. Therefore, membership in $S^{1,1}_{\mu,L}(Q)$ can be characterized by the existence of an oracle returning for each point $y \in Q$ a pair $(f_{L,\mu}(y), g_{L,\mu}(y)) \in \mathbb{R} \times E^*$, necessarily equal to $(f(y), \nabla f(y))$, satisfying

$$\frac{\mu}{2}\|x-y\|_E^2 \;\leq\; f(x) - \big(f_{L,\mu}(y) + \langle g_{L,\mu}(y), x-y\rangle\big) \;\leq\; \frac{L}{2}\|x-y\|_E^2, \quad \forall x \in Q.$$

Our definition of the (δ, L, μ)-oracle consists in introducing a given amount δ of tolerance in this pair of inequalities:

Definition 5.1. Let function f be convex on convex set Q. We say that it is equipped with a first-order (δ, L, μ)-oracle if for any $y \in Q$ we can compute a pair $(f_{\delta,L,\mu}(y), g_{\delta,L,\mu}(y)) \in \mathbb{R} \times E^*$ such that

$$\frac{\mu}{2}\|x-y\|_E^2 \;\leq\; f(x) - \big(f_{\delta,L,\mu}(y) + \langle g_{\delta,L,\mu}(y), x-y\rangle\big) \leq\; \frac{L}{2}\|x-y\|_E^2 + \delta \tag{5.2}$$

for all $x \in Q$ where $\delta \geq 0$ and $L \geq \mu \geq 0$.

This notion of (δ, L, μ)-oracle can be seen as a generalization of the notion of (δ, L)-oracle introduced in the previous chapter. The (δ, L)-oracle has been introduced in order to study the effect of inexact first-order information on the first-order methods designed for an objective function in $F^{1,1}_L(Q)$. We do the same here but for the first-order methods of $S^{1,1}_{\mu,L}(Q)$.

A function f belongs to $S^{1,1}_{\mu,L}(Q)$ if and only it admits a $(0, L, \mu)$-oracle, namely $(f_{0,L,\mu}(y), g_{0,L,\mu}(y)) = (f(y), \nabla f(y))$. However, the class of functions admitting a (δ, L, μ)-oracle is strictly larger, and also includes both nonsmooth functions and functions that are not strongly convex, as we will see in subsection 5.2.5.

5.1.2 Properties

The notions of (δ, L, μ) and (δ, L)-oracles are, of course, strongly related:

- ◇ A (δ, L, μ)-oracle is also a (δ, L)-oracle.
- ◇ A (δ, L)-oracle is a $(\delta, L, 0)$-oracle.

Remark 5.2. Since the notion of (δ, L, μ)-oracle can be seen as a generalization of the notion of (δ, L)-oracle, it could have been possible to present the results of this chapter and of the previous chapter together in a uniform way. However, in order to introduce the new results in a more gradual way, we started in the previous chapter with the case $\mu = 0$ and emphasize in this chapter the case $\mu > 0$. Furthermore, compared to the previous chapter, we restrict ourselves in this chapter to the Euclidean case.

Since a (δ, L, μ)-oracle is also a (δ, L)-oracle, the properties of the (δ, L)-oracle established in the previous chapter (see subsection 4.1.2) are also true for a (δ, L, μ)-oracle. In addition, we would like to highlight here two additional properties of a (δ, L, μ)-oracle that will be useful in the rest of this paper:

• If f admits a (δ, L, μ)-oracle, then cf admits a $(c\delta, cL, c\mu)$-oracle for any value of the constant $c > 0$. If f_i admits a (δ_i, L_i, μ_i)-oracle, $i = 1, 2$, then $f_1 + f_2$ admits a $(\delta_1 + \delta_2, L_1 + L_2, \mu_1 + \mu_2)$-oracle.

•

Theorem 5.1. *If f is endowed with a (δ, L, μ) oracle, we have:*

$$f_{\delta,L,\mu}(\alpha x + (1-\alpha)y) \leq (1-\alpha)f(y) + \alpha f(x) - \frac{\mu}{2}\alpha(1-\alpha)\left\|y - x\right\|_E^2$$

for all $x, y \in E, \alpha \in [0, 1]$ and therefore

$$f(\alpha x + (1-\alpha)y) \leq (1-\alpha)f(y) + \alpha f(x) - \frac{\mu}{2}\alpha(1-\alpha)\left\|y - x\right\|_E^2 + \delta.$$

Proof. Let $x_\alpha = \alpha x + (1-\alpha)y$. We have:

$$\begin{aligned} f(y) &\geq f_{\delta,L,\mu}(x_\alpha) + \langle g_{\delta,L,\mu}(x_\alpha), y - x_\alpha\rangle + \frac{\mu}{2}\left\|x_\alpha - y\right\|_E^2 \\ &= f_{\delta,L,\mu}(x_\alpha) + \alpha\langle g_{\delta,L,\mu}(x_\alpha), y - x\rangle + \frac{\mu}{2}\alpha^2\left\|y - x\right\|_E^2 . \end{aligned}$$

and

$$\begin{aligned} f(x) &\geq f_{\delta,L,\mu}(x_\alpha) + \langle g_{\delta,L,\mu}(x_\alpha), x - x_\alpha\rangle + \frac{\mu}{2}\left\|x - x_\alpha\right\|_E^2 \\ &= f_{\delta,L,\mu}(x_\alpha) + (1-\alpha)\langle g_{\delta,L,\mu}(x_\alpha), x - y\rangle + \frac{\mu}{2}(1-\alpha)^2\left\|y - x\right\|_E^2 . \end{aligned}$$

Adding the first inequality multiplied by $(1-\alpha)$ and the second inequality multiplied by α, we obtain the desired inequality. □

Therefore if we assume that the function f is endowed with a family of $(\delta, L(\delta), \mu(\delta))$-oracles and that

1. $\lim_{\delta \to 0} \mu(\delta) = \overline{\mu} > 0$ then we have:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\overline{\mu}}{2} \alpha (1-\alpha) \|x - y\|_E^2$$

for all $x, y \in E, \alpha \in [0,1]$ and we conclude that f is strongly convex with parameter $\overline{\mu}$

2. $\lim_{\delta \to 0} \mu(\delta) = 0$ then we have:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

for all $x, y \in E, \alpha \in [0,1]$ and we can only conclude that f is convex.

5.2 Examples of (δ, L, μ)-oracle

5.2.1 Strongly convex function with computation at shifted points

Let function $f \in S^{1,1}_{\mu(f), L(f)}(Q)$ be endowed with an oracle providing at each point $y \in Q$, the exact values of the function and its gradient albeit computed at a shifted point $\hat{y}$ different from y.

1. Since f is strongly convex with parameter $\mu(f)$, we have

$$f(x) \geq f(\hat{y}) + \langle \nabla f(\hat{y}), x - \hat{y} \rangle + \frac{\mu(f)}{2} \|x - \hat{y}\|_E^2, \quad \forall x \in Q.$$

Using the convexity of $\|.\|_E^2$, we have $\|x - y\|_E^2 \leq 2\|x - \hat{y}\|_E^2 + 2\|\hat{y} - y\|_E^2$ and therefore

$$f(x) \geq f(\hat{y}) + \langle \nabla f(\hat{y}), x - y \rangle + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \frac{\mu(f)}{4} \|x - y\|_E^2 - \frac{\mu(f)}{2} \|\hat{y} - y\|_E^2. \tag{5.3}$$

2. Since f has a Lipschitz-continuous gradient with constant $L(f)$, we have:

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

Using the convexity of $\|.\|_E^2$, we have $\|x - \hat{y}\|_E^2 \leq 2\|y - \hat{y}\|_E^2 + 2\|x - y\|_E^2$ and therefore

$$f(x) \leq f(\hat{y}) + \langle \nabla f(\hat{y}), x - y \rangle + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + L(f)\|y - \hat{y}\|_E^2 + L(f)\|x - y\|_E^2. \tag{5.4}$$

Letting $\mu = \frac{\mu(f)}{2}$, $L = 2L(f)$ and $\delta = L(f)\|y - \hat{y}\|_E^2 + \frac{\mu(f)}{2}\|y - \hat{y}\|_E^2$, in view of the equations 5.3 and 5.4, we have that

$$(f_{\delta,L,\mu}(y) := f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y}\rangle - \frac{\mu(f)}{2}\|y - \hat{y}\|_E^2, g_{\delta,L,\mu}(y) := \nabla f(\hat{y}))$$

is a (δ, L, μ) oracle for f.

5.2.2 Functions approximated by a smooth strongly convex function

When a function f can be well approximated by a smooth strongly convex function $\bar{f}$, in the sense that their difference is bounded, the exact values of $\bar{f}$ and its gradient provide a (δ, L, μ)-oracle for f. Indeed, assume that there exists a smooth strongly convex function $\bar{f} \in S_{\mu,L}^{1,1}(Q)$ such that $\bar{f}$ is a δ-lower approximation of f on all Q, i.e.

$$0 \leq f(y) - \bar{f}(y) \leq \delta \quad \forall y \in Q.$$

Using the fact that $\bar{f} \in S_{\mu,L}^{1,1}(Q)$, we obtain

$$f(x) \geq \bar{f}(x) \geq \bar{f}(y) + \langle \nabla \bar{f}(y), x - y\rangle + \frac{\mu}{2}\|x - y\|_E^2 \quad \forall x, y \in Q,$$

(using strong convexity of $\bar{f}$), and

$$f(x) \leq \bar{f}(x) + \delta \leq \bar{f}(y) + \langle \nabla \bar{f}(y), x - y\rangle + \frac{L}{2}\|x - y\|_E^2 + \delta \quad \forall x, y \in Q.$$

(using Lipschitz continuity of $\nabla \bar{f}$), which proves that $(\bar{f}(y), \nabla \bar{f}(y))$ is a (δ, L, μ)-oracle for f.

Finally, note that the above result can be readily extended to the case when the δ-lower approximation $\bar{f}$ is not necessarily smooth and strongly convex but is equipped with an inexact (δ', L, μ) oracle: we can then show that the inexact oracle of $\bar{f}$ also constitutes an inexact $(\delta + \delta', L, \mu)$ oracle for f.

5.2.3 Strongly convex function with approximate function value and approximate gradient

Let function $f \in S_{\mu(f),L(f)}^{1,1}(Q)$ be endowed with an oracle that provides us at each point $y \in Q$ with an approximate function value $|f(y) - \tilde{f}_y| \leq \Delta_1$ and an approximate gradient $\left\|\nabla f(y) - \tilde{\nabla} f_y\right\|_E^* \leq \Delta_2$.

Let us prove that this very natural definition of approximate first-order information is a particular case of (δ, L, μ) oracle.

As f is strongly convex with parameter $\mu(f)$, we have

$$\begin{aligned} f(x) &\geq f(y) + \langle \tilde{\nabla} f_y, x - y \rangle \\ &\quad + \langle \nabla f(y) - \tilde{\nabla} f_y, x - y \rangle + \frac{\mu(f)}{2} \|x - y\|_E^2 \\ &\geq f(y) + \langle \tilde{\nabla} f_y, x - y \rangle - \Delta_2 \|x - y\|_E + \frac{\mu(f)}{2} \|x - y\|_E^2 \\ &\geq \tilde{f}_y - \Delta_1 + \langle \tilde{\nabla} f_y, x - y \rangle + \frac{\mu(f)}{4} \|x - y\|_E^2 - \frac{\Delta_2^2}{\mu(f)} \end{aligned}$$

since $\Delta_2 \|x - y\|_E \leq \frac{\Delta_2^2}{\mu(f)} + \frac{\mu(f)}{4} \|x - y\|_E^2$.

As ∇f is Lipschitz-continuous with constant L, we have

$$\begin{aligned} f(x) &\leq f(y) + \langle \tilde{\nabla} f_y, x - y \rangle \\ &\quad + \langle \nabla f(y) - \tilde{\nabla} f_y, x - y \rangle + \frac{L(f)}{2} \|x - y\|_E^2 \\ &\leq f(y) + \langle \tilde{\nabla} f_y, x - y \rangle + \Delta_2 \|x - y\|_E + \frac{L(f)}{2} \|x - y\|_E^2 \\ &\leq \tilde{f}_y + \Delta_1 + \langle \tilde{\nabla} f_y, x - y \rangle + L(f) \|x - y\|_E^2 + \frac{\Delta_2^2}{2L(f)} \end{aligned}$$

since $\Delta_2 \|x - y\|_E \leq \frac{\Delta_2^2}{2L(f)} + \frac{L(f)}{2} \|x - y\|_E^2$.

We conclude that

$$\left(f_{\delta,L,\mu}(y) = \tilde{f}_y - \Delta_1 - \frac{\Delta_2^2}{\mu(f)}, g_{\delta,L,\mu}(y) = \tilde{\nabla} f_y \right)$$

define a (δ, L, μ) oracle for f where $\delta = 2\Delta_1 + \frac{\Delta_2^2}{\mu(f)} + \frac{\Delta_2^2}{2L(f)}$, $\mu = \frac{\mu(f)}{2}$ and $L = 2L(f)$.

In particular, contrarily to the non strongly convex case studied in the previous chapter, we see here that boundedness of Q is not needed for an approximate gradient to fit with our definition of inexact oracle.

5.2.4 Saddle-point functions

Let us now consider objective functions of the form

$$f(x) = \max_{u \in F} \Psi(x, u) = \max_{u \in F} \{G(u) + \langle Au, x \rangle\}$$

where F is a finite-dimensional vector space, endowed with the norm $\|.\|_F$, and $A : F \to E^*$ is a linear operator. We assume that $G : F \to \mathbb{R}$ is

1. Strongly concave with parameter $\mu(G)$ i.e.

$$G(u) \leq G(v) + \langle \nabla G(v), u - v \rangle - \frac{\mu(G)}{2} \|u - v\|_F^2, \quad \forall u, v \in F.$$

2. Smooth with a Lipschitz-continuous gradient with constant $L(G)$

$$G(u) \geq G(v) + \langle \nabla G(v), u - v \rangle - \frac{L(G)}{2} \|u - v\|_F^2, \quad \forall u, v \in F.$$

It is well-known that when $-G \in S^{1,1}_{\mu(G),L(G)}(F)$ then $f \in S^{1,1}_{\mu(f),L(f)}(E)$ where $\mu(f) = \frac{\lambda_{\min}(AA^T)}{L(G)}$ and $L(f) = \frac{\lambda_{\max}(AA^T)}{\mu(G)}$. In particular, the condition numbers of the functions f and G are linked by $Q(f) = \frac{\lambda_{\max}(AA^T)}{\lambda_{\min}(AA^T)} Q(G)$. However, if we want, at a point $z \in E$, to compute the exact first-order information for f, we need to solve the subproblem $\max_{u \in F} \Psi(z, u)$ exactly since $f(z) = \Psi(z, u_z^*)$ and $\nabla f(z) = A u_z^*$ where $u_z^* = \arg\max_{u \in F} \Psi(z, u)$. In practice, we are typically only able to compute an approximate solution $u_z \in F$ of this subproblem. In the following theorem, we give a natural condition under which inexact resolution of the subproblems provides us with a (δ, L, μ)-oracle.

Theorem 5.2. *Assume that G is strongly concave with parameter $\mu(G)$ and smooth with a Lipschitz-continuous gradient with constant $L(G)$. Let $z \in E$ and assume that instead of computing u_z^*, the unique optimal solution of the subproblem $\max_{u \in F} \Psi(z, u)$, we compute $u_z \in F$ such that:*

$$\Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi.$$

Then

$$(f_{\delta,L,\mu}(z) = \Psi(z, u_z) - \xi = G(u_z) + \langle A u_z, z \rangle - \xi, g_{\delta,L,\mu}(z) = A u_z)$$

is a (δ, L, μ) oracle for f with $\delta = 3\xi$, $L = \frac{2\lambda_{\max}(AA^T)}{\mu(G)} = 2L(f)$ and $\mu = \frac{\lambda_{\min}(AA^T)}{2L(G)} = \frac{1}{2}\mu(f)$.

Proof. As $\Psi(z, .)$ has a Lipschitz continuous gradient $\nabla_2 \Psi(z, u) = \nabla G(u) + A^T z$ with constant $L(G)$, we have (see theorem 2.1.5. in [58]):

$$(\left\|\nabla G(u_z) + A^T z\right\|_F^*)^2 \leq 2L(G)(\Psi(z, u_z^*) - \Psi(z, u_z)) \leq 2L(G)\xi. \quad (5.5)$$

◇ The Lipschitz-continuity of ∇G implies (see Lemma 1.2.3 in [58])

$$\begin{aligned} G(u) &\geq G(u_z) + \langle \nabla G(u_z), u - u_z \rangle - \frac{L(G)}{2} \|u - u_z\|_E^2 \\ &= G(u_z) + \langle -A^T z, u - u_z \rangle + \langle \nabla G(u_z) + A^T z, u - u_z \rangle \\ &\quad - \frac{L(G)}{2} \|u - u_z\|_F^2 . \end{aligned}$$

Therefore:

$$\begin{aligned} f(x) &= \max_{u \in F} \{G(u) + \langle Au, x \rangle\} \\ &\geq \max_{u \in F} [G(u_z) + \langle -A^T z, u - u_z \rangle + \langle \nabla G(u_z) + A^T z, u - u_z \rangle \\ &\quad - \frac{L(G)}{2} \|u - u_z\|_F^2 + \langle Au, x \rangle] \\ &= G(u_z) + \langle Au_z, z \rangle + \langle Au_z, x - z \rangle + \max_{u \in F} [\langle A(u - u_z), x - z \rangle \\ &\quad + \langle \nabla G(u_z) + A^T z, u - u_z \rangle - \frac{L(G)}{2} \|u - u_z\|_F^2] \\ &\overset{(5.5)}{\geq} f_{\delta,L,\mu}(z) + \xi + \langle g_{\delta,L,\mu}(z), x - z \rangle + \max_{u \in F} [\langle A(u - u_z), x - z \rangle \\ &\quad - \sqrt{2L(G)\xi} \|u - u_z\|_F - \frac{L(G)}{2} \|u - u_z\|_F^2]. \end{aligned}$$

But $\sqrt{2L(G)\xi} \|u - u_z\|_F \leq \xi + \frac{L(G)}{2} \|u - u_z\|_F^2$ and therefore:

$$\begin{aligned} f(x) &\geq f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z \rangle \\ &\quad + \max_{u \in F} \{\langle A(u - u_z), x - z \rangle - L(G) \|u - u_z\|_F^2\}. \end{aligned}$$

Since

$$\begin{aligned} &\max_{u \in F} \{\langle A(u - u_z), x - z \rangle - L(G) \|u - u_z\|_F^2\} \\ = \; &\frac{1}{4} \frac{\left\|A^T(x - z)\right\|_{F^*}^2}{L(G)} \\ \geq \; &\frac{1}{4} \frac{\lambda_{\min}(AA^T)}{L(G)} \|x - z\|_E^2 \end{aligned}$$

we obtain $f(x) \geq f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z \rangle + \frac{\lambda_{\min}(AA^T)}{4L(G)} \|x - z\|_E^2$.

◇ On the other hand, since G is strongly concave with parameter $\mu(G)$, we

have:

$$\begin{aligned} G(u) &\leq G(u_z^*) + \langle \nabla G(u_z^*), u - z_z^* \rangle - \frac{\mu(G)}{2} \|u - u_z^*\|_F^2 \\ &= G(u_z^*) + \langle -A^T z, u - z_z^* \rangle - \frac{\mu(G)}{2} \|u - u_z^*\|_F^2 \end{aligned}$$

(by definition of u_z^*, we have: $\nabla G(u_z^*) + A^T z = 0$.) Therefore:

$$\begin{aligned} f(x) &= \max_{u \in F} \{G(u) + \langle Au, x \rangle\} \\ &\leq \max_{u \in F} \{G(u_z^*) + \langle -A^T z, u - u_z^* \rangle - \frac{\mu(G)}{2} \|u - u_z^*\|_F^2 + \langle Au, x \rangle\} \\ &= G(u_z^*) + \langle Au_z^*, z \rangle + \langle Au_z^*, x - z \rangle \\ &\quad + \max_{u \in F} \{\langle A(u - u_z^*), x - z \rangle - \frac{\mu(G)}{2} \|u - u_z^*\|_F^2\} \\ &= G(u_z) + (G(u_z^*) - G(u_z)) + \langle Au_z, z \rangle + \langle Au_z, x - z \rangle \\ &\quad + \langle A(u_z^* - u_z), x \rangle + \max_{u \in F} [\langle A(u - u_z), x - z \rangle \\ &\quad + \langle A(u_z - u_z^*), x - z \rangle - \frac{\mu(G)}{2} \|u - u_z^*\|_F^2]. \end{aligned}$$

But $\|u - u_z\|_F^2 \leq 2\|u - u_z^*\|_F^2 + 2\|u_z - u_z^*\|_F^2$ i.e.
$\|u - u_z^*\|_F^2 \geq \frac{1}{2}\|u - u_z\|_F^2 - \|u_z - u_z^*\|_F^2$. Therefore:

$$\begin{aligned} f(x) &\leq G(u_z) + \langle Au_z, z \rangle + \langle Au_z, x - z \rangle + G(u_z^*) - G(u_z) \\ &\quad + \langle A(u_z^* - u_z), z \rangle + \max_{u \in F} [\langle A(u - u_z), x - z \rangle - \frac{\mu(G)}{4} \|u - u_z\|_F^2\} \\ &\quad + \frac{\mu(G)}{2} \|u_z - u_z^*\|_F^2. \end{aligned}$$

Since

1. $\max_{u \in F} \{\langle A(u - u_z), x - z \rangle - \frac{\mu(G)}{4} \|u - u_z\|_F^2\} = \frac{\|A^T(x-z)\|_{F^*}^2}{\mu(G)} \leq \frac{\lambda_{\max}(AA^T)}{\mu(G)} \|x - z\|_E^2$
2. $G(u_z^*) - G(u_z) + \langle A(u_z^* - u_z), z \rangle = \Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi$
3. $\frac{\mu(G)}{2} \|u_z - u_z^*\|_F^2 \leq \Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi$ by strong concavity of $\Psi(z, .)$,

we have:

$$\begin{aligned} f(x) &\leq G(u_z) + \langle Au_z, z \rangle + \langle Au_z, x - z \rangle + \frac{\lambda_{max}(AA^T)}{\mu(G)} \|x - z\|_E^2 + 2\xi \\ &= f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z \rangle + \frac{\lambda_{\max}(AA^T)}{\mu(G)} \|x - z\|_E^2 + 3\xi. \end{aligned}$$

□

5.2.5 Uniformly convex functions with weaker level of smoothness

Let us show that the notion of (δ, L, μ)-oracle can be also useful for solving problems with exact information but where the objective function is not necessarily strongly convex and ∇f not necessarily Lipschitz-continuous. Let function f be subdifferentiable on Q and for each $y \in Q$, denote by $g(y)$ an arbitrary element of the subdifferential $\partial f(y)$.

We assume that

1. f is uniformly convex on Q with convexity parameters $\rho \geq 2$ and $\kappa > 0$ i.e.:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\kappa}{2}\alpha(1-\alpha)\|x-y\|_E^\rho$$

for all $x, y \in Q$ and $\forall \alpha \in [0,1]$. This condition leads to the following inequality:

$$f(x) \geq f(y) + \langle g(y), x-y\rangle + \frac{\kappa}{2}\|x-y\|_E^\rho, \quad \forall x, y \in Q.$$

2. f has an Hölder-continuous (sub)gradient on Q with parameters $\nu \in [0,1]$ an $M < +\infty$ i.e.:

$$\|g(x) - g(y)\|_E^* \leq M\|x-y\|_E^\nu, \quad \forall x, y \in Q.$$

This condition leads to the following inequality:

$$f(x) \leq f(y) + \langle g(y), x-y\rangle + \frac{M}{1+\nu}\|x-y\|_E^{1+\nu}, \quad \forall x, y \in Q.$$

We denote this class of function by:

- $\diamond$ $U_{\kappa,M}^{1,\rho,\nu}(Q)$ when the function is also assumed to be differentiable (it is always the case when $\nu > 0$)
- $\diamond$ $U_{\kappa,M}^{0,\rho,\nu}(Q)$ when the function can be non-differentiable.

Remark 5.3.
- $\diamond$ When $\nu = 0$, the function is typically nonsmooth with bounded variation of the subgradients.
- $\diamond$ When $0 < \nu < 1$, the function is weakly-smooth i.e. with a Hölder-continuous gradient.
- $\diamond$ When $\nu = 1$, the function is smooth with a Lipschitz-continuous gradient.

In particular when $\nu > 0$, the function is necessarily differentiable and we have

$$U_{\kappa,M}^{1,\rho,\nu}(Q) = U_{\kappa,M}^{0,\rho,\nu}(Q)$$

Remark 5.4. When $\rho > 1+\nu$, the class $U^{0,\rho,\nu}_{\kappa,M}(E)$(and therefore also $U^{1,\rho,\nu}_{\kappa,M}(E)$) is empty since

$$\frac{\kappa}{2}t^{\rho} \geq \frac{M}{1+\nu}t^{1+\nu}$$

for sufficiently large $t \geq 0$.

Remark 5.5. $U^{1,2,1}_{\kappa,M}(Q) = S^{1,1}_{\kappa,M}(Q)$.

We will prove in this section that functions $f \in U^{0,\rho,\nu}_{\kappa,M}(Q)$, a (δ, L, μ)-oracle is available for any value of $\delta > 0$ i.e. that we can define quantities $(f_{\delta,L,\mu}(y), g_{\delta,L,\mu}(y))$ satisfying inequalities (5.2).

- ◇ First, we prove that $f(x) \geq f(y) + \langle g(y), x-y\rangle + \frac{\kappa}{2}\|x-y\|^{\rho}_E, \quad \forall x, y \in Q$ implies

  $$f(x) \geq f(y) + \langle g(y), x-y\rangle + \frac{\mu}{2}\|x-y\|^2_E - \delta_1, \quad \forall x, y \in Q$$

  where $\delta_1 > 0$ is arbitrary and for some $\mu > 0$. In order to obtain this implication, we need to find a constant $\mu = \mu(\rho, \kappa, \delta_1)$ such that:

  $$\frac{\mu}{2}\|x-y\|^2_E - \delta_1 \leq \frac{\kappa}{2}\|x-y\|^{\rho}_E, \quad \forall x, y \in Q.$$

  A sufficient condition is $\frac{\mu}{2}t^2 - \delta_1 \leq \frac{\kappa}{2}t^{\rho}$ for all $t \geq 0$. Therefore we will choose $\mu = \min_{t\geq 0}\{\kappa t^{\rho-2} + 2\delta_1 t^{-2}\}$. The optimal solution of this minimization problem is given by $t^* = \left(\frac{4\delta_1}{\kappa(\rho-2)}\right)^{\frac{1}{\rho}}$ and therefore

  $$\mu = \mu(\rho, \kappa, \delta_1) = \rho\left(\frac{1}{\rho-2}\right)^{\frac{\rho-2}{\rho}} \kappa^{\frac{2}{\rho}} \delta_1^{\frac{\rho-2}{\rho}} 2^{1-\frac{4}{\rho}}.$$

  In particular, when $\rho = 2$, we obtain $\mu = \kappa$.

- ◇ Second, in subsection 4.1.3 of Chapter 4, we have proved $f(x) \leq f(y) + \langle g(y), x-y\rangle + \frac{M}{1+\nu}\|x-y\|^{1+\nu}_E$ implies

  $$f(x) \leq f(y) + \langle g(y), x-y\rangle + \frac{L}{2}\|x-y\|^2_E + \delta_2$$

  where δ_2 is arbitrary and $L = M\left(\frac{M}{2\delta_2}\frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}$. In particular when $\nu = 1$, we obtain $L = M$.

We obtain the following theorem:

Theorem 5.3. *Assume that $f \in U^{0,\rho,\nu}_{\kappa,M}(Q)$. Let $0 < \delta_1$ and $0 < \delta_2$ be arbitrary constants and define $\delta = \delta_1 + \delta_2$, $\mu = \rho\left(\frac{1}{\rho-2}\right)^{\frac{\rho-2}{\rho}} \kappa^{\frac{2}{\rho}} \delta_1^{\frac{\rho-2}{\rho}} 2^{1-\frac{4}{\rho}}$ and $L = M\left(\frac{M}{2\delta_2}\frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}$. Then*

$$(f_{\delta,L,\mu}(y) = f(y) - \delta_1, g_{\delta,L,\mu}(y) = g(y) \in \partial f(y))$$

defines a (δ, L, μ) oracle for f.

5.3 Primal Gradient Method with (δ, L, μ)-oracle

Let us now study the behavior of first-order methods, initially developed for smooth strongly convex problems, but used here with a (δ, L, μ)-oracle.

We start with the Primal Gradient Method. One important property of this method is that it does not use the strongly convex parameter μ explicitly in the scheme. The Primal Gradient Method for strongly convex problems looks exactly the same as in the convex case. Therefore, the Primal Gradient Method when used with a (δ, L, μ) oracle looks exactly the same that when used with a (δ, L)-oracle.

Algorithm 16 Primal Gradient Method (PGM) with (δ, L, μ) oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute $x_{k+1} = T_L(x_k, g_{\delta,L,\mu}(x_k))$
5: **end for**

Even if the scheme is the same, the fact that we use a (δ, L, μ)-oracle instead of a (δ, L)-oracle can accelerate significantly the convergence rate:

Theorem 5.4. *Assume that f is endowed with a (δ, L, μ)-oracle with $\mu > 0$, then the sequence $y_k = arg\min_{x_1,\ldots,x_k} f(x_i)$, generated by the Primal Gradient Method satisfies*

$$f(y_k) - f^* \leq \frac{LR^2}{2}\exp\left(-k\frac{\mu}{L}\right) + \delta.$$

Proof. Denote $r_k = \|x_k - x^*\|_E$ and $f_k = f_{\delta,L,\mu}(x_k)$, $g_k = g_{\delta,L,\mu}(x_k)$. We have

$$r^2_{k+1} = \|x_{k+1} - x^*\|^2_E = r^2_k + 2\langle B(x_{k+1} - x_k), x_{k+1} - x^*\rangle - \|x_{k+1} - x_k\|^2_E \tag{5.6}$$

Using the optimality condition of the problem defining x_{k+1}:

$$\langle g_k + LB(x_{k+1} - x_k), x - x_{k+1} \rangle \geq 0 \quad \forall x \in Q$$

we have

$$\langle B(x_{k+1} - x_k), x_{k+1} - x^* \rangle \leq \frac{1}{L} \langle g_k, x^* - x_{k+1} \rangle.$$

We obtain

$$\begin{aligned}
r_{k+1}^2 &\leq r_k^2 + \frac{2}{L}\langle g_k, x^* - x_{k+1}\rangle - \|x_{k+1} - x_k\|_E^2 \\
&= r_k^2 + \frac{2}{L}\langle g_k, x^* - x_k\rangle - \frac{2}{L}\left[\langle g_k, x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_E^2\right] \\
&\overset{(5.2)}{\leq} r_k^2 + \frac{2}{L}\langle g_k, x^* - x_k\rangle - \frac{2}{L}\left[f(x_{k+1}) - f_k - \delta\right] \\
&\overset{(5.2)}{\leq} r_k^2 + \frac{2}{L}\left[f(x^*) - f_k - \frac{\mu}{2}\|x_k - x^*\|_E^2\right] - \frac{2}{L}\left[f(x_{k+1}) - f_k - \delta\right] \\
&= \left(1 - \frac{\mu}{L}\right) r_{k+1}^2 + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
r_{k+1}^2 &\leq \left(1 - \frac{\mu}{L}\right) r_k^2 + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right] \\
&\leq \left(1 - \frac{\mu}{L}\right)\left(\left(1 - \frac{\mu}{L}\right) r_{k-1}^2 + \frac{2}{L}\left[f(x^*) - f(x_k) + \delta\right]\right) \\
&\quad + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right] \\
&\leq \left(1 - \frac{\mu}{L}\right)^k r_0^2 + \frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i (f(x^*) - f(x_{k+1-i}) + \delta).
\end{aligned}$$

and we obtain

$$\frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i (f(x_{k+1-i}) - f(x^*)) \leq \left(1 - \frac{\mu}{L}\right)^{k+1} r_0^2 + \frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i \delta.$$

Therefore, using the definition of y_{k+1} and the fact that $\frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i = \frac{2}{\mu}\left(1 - (1 - \frac{\mu}{L})^{k+1}\right)$ we conclude that

$$\begin{aligned}
f(y_{k+1}) - f^* &\leq \frac{\mu}{2}\frac{\left(1 - \frac{\mu}{L}\right)^{k+1}}{1 - \left(1 - \frac{\mu}{L}\right)^{k+1}} r_0^2 + \delta \\
&\leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{k+1} r_0^2 + \delta \\
&\leq \frac{L r_0^2}{2}\exp\left(-(k+1)\frac{\mu}{L}\right) + \delta.
\end{aligned}$$

□

Remark 5.6. When $\delta = 0$, we retrieve the well-known behavior of the Primal Gradient Method in the strongly convex case, with a complexity of order $\Theta\left(\frac{L}{\mu}\ln\left(\frac{LR^2}{\epsilon}\right)\right)$.

Remark 5.7. As a (δ, L, μ) oracle is also a (δ, L)-oracle and as the parameter μ is not used during the scheme, the upper-bound

$$f(y_k) - f^* \leq \frac{LR^2}{2k} + \delta \tag{5.7}$$

obtained in the previous chapter, is still available. Therefore, the sequence y_k generated by the primal gradient method actually satisfies

$$f(y_k) - f^* \leq \frac{LR^2}{2}\min\left(\frac{1}{k}, \exp\left(-k\frac{\mu}{L}\right)\right) + \delta.$$

In conclusion, when we apply the primal gradient method to a function endowed with a (δ, L, μ) oracle, there is no error accumulation, and the upper bound for the objective function accuracy decreases with k and asymptotically tends to δ. If we want an accuracy of ϵ for the objective function, we need to perform a number of iterations k such that

$$k = \min\left(\Theta\left(\frac{LR^2}{\epsilon}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$$

with an oracle accuracy $\delta = \Theta(\epsilon)$. As in the non strongly convex case, studied in the previous chapter, the PGM does not suffer from errors accumulation.

5.4 Dual Gradient Method with (δ, L, μ)-oracle

Let us now consider the Dual Gradient Method. This method has been introduced in [62] for smooth convex problems with exact oracle. In the previous chapter, we have studied its behavior when used with a (δ, L)-oracle.

In the strongly convex case, it is necessary to modify slightly the method in order to take advantage of the strong convexity. We propose here such modification and study the behavior of this modified method when used with a (δ, L, μ) oracle.

Let $\{\alpha_k\}_{k\geq 0}$ be a sequence of positive reals such that:

$$\alpha_0 = \frac{L}{L-\mu} \tag{5.8}$$

$$(L-\mu)\alpha_{k+1} = A_k\mu + L \tag{5.9}$$

where $A_k = \sum_{i=0}^{k} \alpha_i$.

Algorithm 17 Dual Gradient Method (DGM) with (δ, L, μ) oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute

$$w_k = T_L(x_k, g_{\delta,L,\mu}(x_k)) \tag{5.10}$$

5: Compute

$$x_{k+1} = \arg\min_{x\in Q} \left[\sum_{i=0}^{k} \alpha_i [\langle g_{\delta,L,\mu}(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_E^2] + \frac{L}{2}\|x - x_0\|_E^2 \right].$$

6: **end for**

Lemma 5.5. *For any $k \geq 0$ we have*

$$\sum_{i=0}^{k} \alpha_i [f(w_i) - f^*] \quad \leq \quad \tfrac{L}{2}\|x_0 - x^*\|_E^2 + \sum_{i=0}^{k} \alpha_i \delta \tag{5.11}$$

Proof. For $k \geq 0$, denote $f_k = f_{\delta,L,\mu}(x_k)$, $g_k = g_{\delta,L,\mu}(x_k)$, and $\psi_k^* = \min\limits_{x\in Q} \psi_k(x)$ where

$$\psi_k(x) \quad = \quad \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, x - x_i\rangle + \tfrac{\mu}{2}\|x - x_i\|_E^2] + \tfrac{L}{2}\|x - x_0\|_E^2.$$

In view of the first inequality in (5.2), we have for all $x \in Q$

$$\psi_k^* \quad \leq \quad \psi_k(x) \ \leq \ \textstyle\sum_{i=0}^{k} \alpha_i f(x) + \frac{L}{2}\|x - x_0\|_E^2 . \tag{5.12}$$

Let us prove that $\psi_k^* \geq \sum\limits_{i=0}^{k} \alpha_i [f(w_i) - \delta], \quad \forall k \geq 0.$

Indeed, this inequality is valid for $k = 0$:

$$\begin{aligned}
\alpha_0 f(w_0) &\overset{(5.2)}{\leq} \alpha_0 [f_0 + \langle g_0, w_0 - x_0\rangle + \frac{L}{2}\|w_0 - x_0\|_E^2 + \delta] \\
&\overset{(5.10)}{=} \min_{y\in Q} \alpha_0 [f_0 + \langle g_0, y - x_0\rangle + \frac{L}{2}\|y - x_0\|_E^2] + \alpha_0\delta \\
&\leq \min_{y\in Q} \{\alpha_0 [f_0 + \langle g_0, y - x_0\rangle + \frac{\mu}{2}\|y - x_0\|_E^2] + \frac{L}{2}\|y - x_0\|_E^2\} + \alpha_0\delta \\
&= \psi_0^* + \alpha_0\delta.
\end{aligned}$$

Assume it is valid for some $k \geq 1$. Since $\Psi_k(x)$ is strongly convex with parameter $\sum_{i=0}^{k} \alpha_i \mu + L = A_k \mu + L$, we have:

$$\psi_k(x) \quad \geq \quad \psi_k^* + \tfrac{A_k \mu + L}{2} \|x - x_{k+1}\|_E^2, \quad x \in Q$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &= \min_{x \in Q} \left\{ \psi_k(x) + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \right\} \\
&\geq \psi_k^* + \min_{x \in Q} \{ \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\
&\quad + \frac{A_k \mu + L}{2} \|x - x_{k+1}\|_E^2 \} \\
&\geq \psi_k^* + \alpha_{k+1} \min_{x \in Q} \{ f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{L}{2} \|x - x_{k+1}\|_E^2 \}
\end{aligned}$$

since $\alpha_{k+1}\mu + A_k \mu + L = L\alpha_{k+1}$.
And we obtain finally :

$$\psi_{k+1}^* \overset{(5.10),(5.2)}{\geq} \psi_k^* + \alpha_{k+1}(f(w_{k+1}) - \delta).$$

Hence, using our inductive assumption, we have proved that $\psi_k^* \geq \sum_{i=0}^{k} \alpha_i [f(w_i) - \delta]$ for all $k \geq 0$. To conclude, we combine this fact with inequality (5.12) for $x = x^*$. □

Defining now the approximate solution as $y_k = \arg\min_{i=0,\ldots,k} f(w_i)$ or $y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{\sum_{i=0}^{k} \alpha_i}$ we obtain:

$$f(y_k) - f^* \leq \frac{L \|x_0 - x^*\|_E^2}{2 \sum_{i=0}^{k} \alpha_i} + \delta = \frac{LR^2}{2A_k} + \delta. \tag{5.13}$$

It remains therefore to obtain a lower bound for A_k:

Lemma 5.6. *The sequence $\{A_k\}_{k \geq 0}$ defined by the recurrence (5.8) and (5.9) satisfies*

$$A_k = \sum_{i=1}^{k+1} \left(\frac{L}{L - \mu} \right)^i$$

and therefore

1. $A_k = k + 1$, $\forall k \geq 0$ *if* $\mu = 0$
2. $A_k = \frac{L}{\mu} \left(\left(\frac{L}{L-\mu} \right)^{k+1} - 1 \right)$, $\forall k \geq 0$ *if* $\mu > 0$.

Proof. We have

$$(L-\mu)\alpha_{k+1} = A_k\mu + L \Leftrightarrow (L-\mu)A_{k+1} = L(A_k+1)$$

and therefore

$$A_{k+1} = \frac{L}{L-\mu}(A_k+1).$$

As $A_0 = \alpha_0 = \frac{L}{L-\mu}$, we conclude that

$$A_k = \sum_{i=1}^{k+1}\left(\frac{L}{L-\mu}\right)^i.$$

□

We obtain finally the following theorem:

Theorem 5.7. *The dual gradient method applied to a function f endowed with a (δ, L, μ)-oracle generates a sequence $\{w_k\}_{k\geq 0}$ such that $y_k = \arg\min_{i=0,\dots,k} f(w_i)$ and $y_k = \frac{\sum_{i=0}^k \alpha_i w_i}{\sum_{i=0}^k \alpha_i}$ satisfy*

$$f(y_k) - f^* \leq \frac{LR^2}{2(k+1)} + \delta, \text{ if } \mu = 0$$

and

$$f(y_k) - f^* \leq \frac{\mu R^2}{2\left(\left(\frac{L}{L-\mu}\right)^{k+1} - 1\right)} + \delta \leq \frac{LR^2}{2}\exp\left(-(k+1)\frac{\mu}{L}\right) + \delta$$

if $\mu > 0$.

Remark 5.8. When $\mu = 0$, we have $\alpha_i = 1\ \forall i \geq 0$ and this method corresponds to the dual gradient method introduced in [62] and for which the behavior when used with a $(\delta, L) = (\delta, L, 0)$ has been already established in Chapter 4.

Remark 5.9. In the case $\mu > 0$, the sequence $A_k(\mu) = A_k$ satisfies the recurrence

$$A_{k+1}(\mu) = \frac{L}{L-\mu}A_k(\mu) + \frac{L}{L-\mu}$$

and in the case $\mu = 0$, the sequence $A_k(0)$ satisfies the recurrence

$$A_{k+1}(0) = A_k(0) + 1.$$

Clearly $A_k(\mu) \geq A_k(0)$ for all $k \geq 1$ and

$$f(y_k) - f^* \leq \frac{LR^2}{2A_k(\mu)} + \delta \leq \frac{LR^2}{2A_k(0)} + \delta.$$

Therefore the upper-bound

$$f(y_k) - f^* \leq \frac{LR^2}{2(k+1)} + \delta$$

is also available in the case $\mu > 0$ and we have:

$$f(y_k) - f^* \leq \min\left(\frac{LR^2}{2(k+1)}, LR^2 \exp\left(-k\frac{\mu}{L}\right)\right) + \delta.$$

Remark 5.10. When $\delta = 0$, the availability of a $(0, L, \mu)$ oracle for a function f means simply that $f \in S^{1,1}_{\mu,L}(Q)$. To the best of our knowledge, it is the first time that the dual gradient method is adapted to the strongly convex case.

Since we obtain the same convergence results for both primal and dual gradient methods, we will refer to both as Gradient Methods (GM) in the rest of this chapter.

5.5 Fast Gradient Method with (δ, L, μ)-oracle

The fast gradient method (at least the version that we consider in this thesis) has been introduced in [59]. In the previous chapter, we have studied the behavior of this scheme when used with a (δ, L)-oracle instead of the exact one. In this section, we adapt this fast-gradient method to the strongly convex case and we apply this scheme to a convex function f endowed with a (δ, L, μ)-oracle.

5.5.1 The method

Let $\{\alpha_k\}_{k\geq 0}$ be a sequence of reals such that

$$L + \mu A_k = \frac{L\alpha_{k+1}^2}{A_{k+1}}, \quad \alpha_0 = 1 \tag{5.14}$$

where $A_k = \sum_{i=0}^{k} \alpha_i$. Define $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$, $k \geq 0$. The condition on the sequence $\{\alpha_k\}_{k=0}^{\infty}$ is equivalent with

$$\frac{L + \mu A_k}{A_{k+1}} = L\tau_k^2. \tag{5.15}$$

For a particular setup choice, the FGM used with a (δ, L, μ)-oracle looks as follows

Algorithm 18 Fast Gradient Method (FGM) with (δ, L, μ) oracle

1: Choose $x_0 = \min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute

$$y_k = T_L(x_k, g_{\delta,L,\mu}(x_k)) \tag{5.16}$$

5: Compute

$$z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [\langle g_{\delta,L,\mu}(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2]\}$$

6: Define $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
7: **end for**

5.5.2 Convergence rate

Denote

$$\psi_k^* = \min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f_{\delta,L,\mu}(x_i) + \langle g_{\delta,L,\mu}(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2]\}.$$

Lemma 5.8. *For all $k \geq 0$, we have $A_k f(y_k) \leq \psi_k^* + E_k$ with $E_k = \sum_{i=0}^{k} A_i \delta$.*

Proof. Denote $f_k = f_{\delta,L,\mu}(x_k)$, and $g_k = g_{\delta,L,\mu}(x_k)$. For $k = 0$, we have

$$\begin{aligned} \psi_0^* &= \min_{x \in Q} \left\{ Ld(x) + \alpha_0 [f_0 + \langle g_0, x - x_0 \rangle + \tfrac{\mu}{2} \|x - x_0\|_E^2] \right\} \\ &\overset{(2.8)}{\geq} \min_{x \in Q} \left\{ f_0 + \langle g_0, x - x_0 \rangle + \tfrac{L}{2} \|x - x_0\|^2 \right\} \overset{(5.2)}{\geq} [f(y_0) - \delta]. \end{aligned}$$

since $\alpha_0 = 1$.
Assume now that the statement of the lemma is true for some $k \geq 0$. Optimality condition for the optimization problem defining z_k implies

$$\langle \nabla Ld(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i g_i + \sum_{i=0}^{k} \alpha_i \mu B(z_k - x_i), x - z_k \rangle \quad \geq \quad 0, \quad \forall x \in Q.$$

Hence, in view of strong convexity of d,

$$\begin{aligned} Ld(x) &\geq Ld(z_k) + \langle L\nabla d(z_k), x - z_k\rangle + \frac{L}{2}\|x - z_k\|_E^2 \\ &\geq Ld(z_k) + \sum_{i=0}^{k} \alpha_i \langle g_i, z_k - x\rangle \\ &\quad + \sum_{i=0}^{k} \alpha_i \mu \langle B(z_k - x_i), z_k - x\rangle + \frac{L}{2}\|x - z_k\|_E^2 . \end{aligned}$$

Thus, we have for all $x \in Q$:

$$\begin{aligned} & Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_E^2] \\ \geq\ & Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i\rangle] + \frac{L}{2}\|x - z_k\|_E^2 \\ & + \sum_{i=0}^{k} \alpha_i \mu \langle B(z_k - x_i), z_k - x\rangle + \sum_{i=0}^{k} \frac{\alpha_i \mu}{2}\|x - x_i\|_E^2 \\ & + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2]. \end{aligned}$$

But:

$$\langle B(z_k - x_i), z_k - x\rangle = \frac{1}{2}\|z_k - x_i\|_E^2 + \frac{1}{2}\|z_k - x\|_E^2 - \frac{1}{2}\|x - x_i\|_E^2$$

and we obtain:

$$Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_E^2]$$

$$\geq Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i\rangle + \frac{\mu}{2}\|z_k - x_i\|_E^2]$$

$$+ \frac{L + A_k\mu}{2}\|z_k - x\|_E^2 + + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2].$$

which implies:

$$\begin{aligned} \psi_{k+1}^* &\geq \psi_k^* + \min_{x\in Q}\{\tfrac{L+\mu A_k}{2}\|x - z_k\|_E^2 + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle \\ &\quad + \tfrac{\mu}{2}\|x - x_{k+1}\|_E^2]\}. \end{aligned}$$

On the other hand, using our recurrence assumption $A_k f(y_k) \leq \psi_k^* + E_k$, we have

$$\begin{aligned}
& \psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2] \\
\geq\ & A_k f(y_k) - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2] \\
\overset{(5.2)}{\geq}\ & A_k[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle + \frac{\mu}{2}\|y_k - x_{k+1}\|_E^2] - E_k \\
& + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2] \\
=\ & A_{k+1} f_{k+1} + \langle g_{k+1}, A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle \\
& - E_k + \frac{A_k \mu}{2}\|y_k - x_{k+1}\|_E^2 + \frac{\alpha_{k+1}\mu}{2}\|x - x_{k+1}\|_E^2 .
\end{aligned}$$

Taking into account that

$$\begin{aligned}
& A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1}) \\
=\ & A_k\tau_k(y_k - z_k) + \alpha_{k+1}x - \alpha_{k+1}\tau_k z_k - \alpha_{k+1}(1-\tau_k)y_k \ =\ \alpha_{k+1}(x - z_k),
\end{aligned}$$

we obtain

$$\begin{aligned}
& \psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{\mu}{2}\|x - x_{k+1}\|_E^2] \\
\geq\ & A_{k+1} f_{k+1} + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle - E_k.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* \quad \geq\ & A_{k+1} f_{k+1} - E_k + \min_{x \in Q}\{\tfrac{L+\mu A_k}{2}\|x - z_k\|_E^2 + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle\} \\
=\ & A_{k+1}\left[f_{k+1} + \min_{x \in Q}\{\tfrac{L+\mu A_k}{2A_{k+1}}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}\right] - E_k \\
\overset{(5.15)}{=}\ & A_{k+1}\left[f_{k+1} + \min\nolimits_{x \in Q}\{\tfrac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}\right] - E_k.
\end{aligned}$$

For $x \in Q$, define $y = \tau_k x + (1 - \tau_k)y_k$. Since $y - x_{k+1} = \tau_k(x - z_k)$, we obtain

$$\begin{aligned}
& \min_{x \in Q}\left\{\tfrac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\right\} \\
=\ & \min_{y}\left\{\tfrac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle :\ y \in \tau_k Q + (1-\tau_k)y_k\right\} \\
\geq\ & \min_{y \in Q}\left\{\tfrac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle\right\}.
\end{aligned} \tag{5.17}$$

Therefore,we have

$$\begin{aligned}\psi^*_{k+1} &\geq A_{k+1}\left[f_{k+1} + \min_{y\in Q}\{\tfrac{L}{2}\|y - x_{k+1}\|^2 + \langle g_{k+1}, y - x_{k+1}\rangle\}\right] - E_k \\ &\overset{(5.16),(5.2)}{\geq} A_{k+1}f(y_{k+1}) - E_k - A_{k+1}\delta,\end{aligned}$$

and we get $A_{k+1}f(y_{k+1}) \leq \psi_{k+1} + E_{k+1}$ with $E_{k+1} = E_k + A_{k+1}\delta$. □

As a direct consequence of this lemma, we obtain

Theorem 5.9. *For all $k \geq 0$, we have $f(y_k) - f^* \leq \frac{1}{A_k}\left(Ld(x^*) + \sum_{i=0}^k A_i\delta\right)$.*

Proof. Denote $f_i = f_{\delta,L,\mu}(x_i)$, and $g_i = g_{\delta,L,\mu}(x_i)$. Then

$$\begin{aligned}\psi^*_k &= \min_{x\in Q}\left\{Ld(x) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x - x_i\rangle + \tfrac{\mu}{2}\|x - x_i\|^2_E]\right\} \\ &\leq Ld(x^*) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x^* - x_i\rangle + \tfrac{\mu}{2}\|x^* - x_i\|^2_E] \\ &\overset{(5.2)}{\leq} Ld(x^*) + A_k f(x^*).\end{aligned}$$

The proof now simply follows from the recurrence established in Lemma 5.8. □

It remains to estimate A_k and $\frac{\sum_{i=0}^k A_i}{A_k}$. More precisely, in order to obtain an explicit upper-bound for the convergence rate of this method, we need

1. A lower-bound for A_k
2. An upper-bound for $\frac{\sum_{i=0}^k A_i}{A_k}$.

Concerning the lower bound for A_k, we have the following result

Lemma 5.10. *The sequence $\{A_k\}_{k\geq 0}$ defined by the recurrence (5.14) satisfies*

$$A_k \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k} \quad \forall k \geq 0.$$

Proof. We have:

$$\begin{aligned}\mu A_k A_{k+1} &\leq L(A_{k+1} - A_k)^2 = L(A^{1/2}_{k+1} - A^{1/2}_k)^2(A^{1/2}_{k+1} + A^{1/2}_k)^2 \\ &\leq 4LA_{k+1}(A^{1/2}_{k+1} - A^{1/2}_k)^2.\end{aligned}$$

Therefore $\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)A^{1/2}_k \leq A^{1/2}_{k+1}$ which implies $A_k \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}$. □

Concerning $\frac{\sum_{i=0}^{k} A_i}{A_k}$, the cumulative effect of the successive oracle errors, we begin with the following uniform upper-bound:

Lemma 5.11. *The sequence $\{A_k\}_{k\geq 0}$ defined by the recurrence (5.14) satisfies*

$$\frac{\sum_{i=0}^{k} A_i}{A_k} \leq 1 + \frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}} + \sqrt{\frac{\mu}{L}+4}} \leq 1 + \sqrt{\frac{L}{\mu}} \quad \forall k \geq 0.$$

Proof. We first note that A_k satisfies the following recurrence equation:

$$A_{k+1}^2 - \left(1 + A_k\left(\frac{\mu}{L}\right) + 2A_k\right) A_{k+1} + A_k^2 = 0$$

or equivalently:

$$A_{k+1} = \frac{\left(1 + A_k\left(\frac{\mu}{L}\right) + 2A_k\right) + \sqrt{\left(1 + A_k\left(\frac{\mu}{L}\right) + 2A_k\right)^2 - 4A_k^2}}{2}.$$

For our analysis, we consider also the sequence defined by the recurrence $\mu\tilde{A}_k = \frac{L(\tilde{A}_{k+1}-\tilde{A}_k)^2}{\tilde{A}_{k+1}}$ or equivalently $\tilde{A}_{k+1} = \tilde{A}_k\left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2} + 1\right)$. We have clearly $\frac{A_{k+1}}{A_k} \geq \frac{\tilde{A}_{k+1}}{\tilde{A}_k} \quad \forall k \geq 0$ and therefore $\frac{A_k}{A_i} \geq \frac{\tilde{A}_k}{\tilde{A}_i} \quad \forall i < k, \quad \forall k \geq 1$. We conclude that:

$$\frac{\sum_{i=0}^{k} A_i}{A_k} \leq \frac{\sum_{i=0}^{k} \tilde{A}_i}{\tilde{A}_k}.$$

On the other hand, if we assume that $\tilde{A}_0 = A_0 = 1$, we have $\tilde{A}_k = C^k$ where $C = \left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2} + 1\right)$. Therefore, we have $\sum_{i=0}^{k} \tilde{A}_i = \sum_{i=0}^{k} C^i = \frac{C^{k+1}-1}{C-1}$ and

$$\begin{aligned}
\frac{\sum_{i=0}^{k} \tilde{A}_i}{\tilde{A}_k} &= \frac{C^{k+1}-1}{C-1}\frac{1}{C^k} = \frac{C^{k+1}-1}{C^{k+1}-C^k} \\
&\leq \frac{C}{C-1} = \frac{\frac{\mu}{L} + \sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4} + 2}{\frac{\mu}{L} + \sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}} = 1 + \frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}} + \sqrt{\frac{\mu}{L}+4}} \\
&\leq 1 + \sqrt{\frac{L}{\mu}}.
\end{aligned}$$

We conclude that:

$$\frac{\sum_{i=0}^{k} A_i}{A_k} \leq \frac{\sum_{i=0}^{k} \tilde{A}_i}{\tilde{A}_k} \leq 1 + \frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}} + \sqrt{\frac{\mu}{L}+4}} \leq 1 + \sqrt{\frac{L}{\mu}}.$$

□

A pessimistic but also an optimistic interpretation of this result can be done:

- ◇ The FGM is worse than the GM concerning the effect of the oracle errors. Contrarily to the gradient method for which the oracle accuracy δ can be chosen of the same level that the desired final accuracy, the fast-gradient method suffers from an increase of error. The cumulative error (in the convergence rate for $f(y_k)-f^*$) coming from the successive oracle errors is bigger than each individual oracle error δ. This bad phenomenon does not come from our analysis but is a unavoidable problem of any fast first-order method for smooth strongly convex problems as we will see in Theorem 5.15. More precisely, for any optimal method in smooth strongly convex optimization, the total effect on the convergence rate of the successive oracle errors cannot be bounded by a uniform quantity (i.e. independent of k) having a better dependence in the condition number than $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$.
- ◇ When $\mu > 0$, the FGM does not suffer from an unbounded accumulation of errors. When $\mu = 0$, i.e. when the function is endowed with a (δ, L)-oracle, we have established in the previous chapter that the fast gradient method suffers from an accumulation of oracle errors with rate $\Theta(k\delta)$, making the method asymptotically divergent. When $\mu > 0$, we can bound the total effect of the oracle errors by a quantity of order $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ that does not depend of k. This method is not divergent, the error on the function value converges to a limit smaller than $\left(1+\sqrt{\frac{L}{\mu}}\right)\delta$.

Of course, the cumulative effect of the oracle errors does not reach the level $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ from the very first iterations. In fact, we will prove now that the effect of the oracle errors is always less undesirable in the case $\mu > 0$ than in the case $\mu = 0$. We can bound $\frac{\sum_{i=0}^{k} A_i}{A_k}$ by an uniform quantity (impossible when $\mu = 0$) but also by the quantity of order $\Theta(k\Theta)$ available in the case $\mu = 0$:

Lemma 5.12. *Let $\mu > 0$. The sequences $\{A_k(\mu)\}_{k\geq 0}$ and $\{A_k(0)\}_{k\geq 0}$ defined by the recurrences:*

$$L + \mu A_k(\mu) = \frac{L(A_{k+1}(\mu) - A_k(\mu))^2}{A_{k+1}(\mu)}, \quad A_0(\mu) = 1$$

$$A_{k+1}(0) = (A_{k+1}(0) - A_k(0))^2, \quad A_0(0) = 1$$

satisfy:

$$\frac{\sum_{i=0}^{k} A_i(\mu)}{A_k(\mu)} \leq \frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)}.$$

In order to prove this result, we first establish the following lemma:

Lemma 5.13. *For all $\mu > 0$, we have: $\frac{1}{A_k(\mu)} + \frac{\mu}{L} \geq \frac{1}{A_k(0)}$ i.e.: $A_k(0) \geq \frac{LA_k(\mu)}{L+A_k(\mu)\mu}$.*

Proof. $\diamond$ It is true for $k = 0$. Indeed as $A_0(0) = A_0(\mu) = 1$, we have

$$\frac{1}{A_0(\mu)} + \frac{\mu}{L} \geq \frac{1}{A_0(0)}.$$

$\diamond$ Assume it is true for $k \geq 0$. We have:

$$\begin{aligned}
A_{k+1}(0) &= \frac{1 + 2A_k(0) + \sqrt{(1 + 2A_k(0))^2 - 4A_k(0)^2}}{2} \\
&= \frac{1 + 2A_k(0) + \sqrt{1 + 4A_k(0)}}{2} \\
&\geq \frac{1 + \frac{2LA_k(\mu)}{L+\mu A_k(\mu)} + \sqrt{1 + \frac{4LA_k(\mu)}{L+\mu A_k(\mu)}}}{2} \\
&= \frac{L + \mu A_k(\mu) + 2LA_k(\mu)}{2(L + \mu A_k(\mu))} \\
&\quad + \frac{\sqrt{(L + A_k(\mu)\mu)^2 + 4LA_k(\mu)(L + A_k(\mu)\mu)}}{2(L + \mu A_k(\mu))} \\
&= \frac{L + (\mu)\, A_k(\mu) + 2LA_k(\mu)}{2(L + \mu A_k(\mu))} \\
&\quad + \frac{\sqrt{(L + A_k(\mu)\mu + 2L^2 A_k(\mu))^2 - 4L^2 A_k(\mu)^2}}{2(L + A_k(\mu)\mu)} \\
&= \frac{LA_{k+1}(\mu)}{L + \mu A_k(\mu)} \\
&\geq \frac{LA_{k+1}(\mu)}{L + \mu A_{k+1}(\mu)}.
\end{aligned}$$

since $A_{k+1}(\mu) \geq A_k(\mu)$.

□

We are now able to give the proof of the Lemma 5.12:

Proof. We have:

$$\frac{A_{k+1}(0)}{A_k(0)} = \frac{\frac{1}{A_k(0)} + 2 + \sqrt{\left(\frac{1}{A_k(0)} + 2\right)^2 - 4}}{2}$$

and

$$\frac{A_{k+1}(\mu)}{A_k(\mu)} = \frac{\frac{1}{A_k(\mu)} + \frac{\mu}{L} + 2 + \sqrt{\left(\frac{1}{A_k(\mu)} + \frac{\mu}{L} + 2\right)^2 - 4}}{2}.$$

Therefore as

$$\frac{1}{A_k(0)} \leq \frac{1}{A_k(\mu)} + \frac{\mu}{L}$$

using Lemma 5.13, we have:

$$\frac{A_{k+1}(0)}{A_k(0)} \leq \frac{A_{k+1}(\mu)}{A_k(\mu)}, \quad \forall k \geq 0.$$

As a consequence, we obtain:

$$\frac{A_k(0)}{A_i(0)} \leq \frac{A_k(\mu)}{A_i(\mu)}, \quad \forall 0 \geq i < k$$

and therefore

$$\frac{\sum_{i=0}^{k} A_i(\mu)}{A_k(\mu)} \leq \frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)}.$$

□

Remark 5.11. The behavior of the FGM in the case $\mu > 0$ is never worse than in the case $\mu = 0$. This is true for the rate of error accumulation, as we have seen in the theorem 5.12, but also for the convergence rate in the exact case (i.e. the first term of the convergence rate, the term that does not depend on δ). Indeed, since $A_k(\mu) \geq A_k(0)$ for all $k \geq 0$, the first term in the convergence rate of the FGM i.e. $\frac{Ld(x^*)}{A_k(\mu)}$ can be bounded by $\frac{Ld(x^*)}{\left(1+\frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}}$ as we have seen in the theorem 5.10 but also by the upper-bound $\frac{Ld(x^*)}{A_k(0)} = \Theta\left(\frac{LR^2}{k^2}\right)$ available in the case $\mu = 0$ (see [24]).

Remark 5.12. The fast gradient method presented here is compatible with the case $\mu = 0$ but is not completely equivalent in the choice of the sequence $\{\alpha_i\}_{i\geq 0}$ with the version analyzed in Chapter 4. However, the two methods present the same rate of convergence and the same rate of errors accumulation. More precisely, in the FGM presented here, we have

$$\begin{aligned} A_{k+1}(0) &= (A_{k+1}(0) - A_k(0))^2 \\ &= (A_{k+1}(0)^{1/2} - A_k(0)^{1/2})^2 (A_{k+1}(0)^{1/2} + A_k(0)^{1/2})^2 \\ &\leq 4A_{k+1}(0)(A_{k+1}(0)^{1/2} - A_k(0)^{1/2})^2 \end{aligned}$$

and therefore:

$$A_{k+1}(0)^{1/2} \geq \frac{1}{2} + A_k(0)^{1/2}$$

which implies:

$$A_{k+1}(0) \geq \frac{(k+1)^2}{4}.$$

Furthermore, it is possible to show that

$$\frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)} \leq \frac{1}{3}k + 2.4.$$

We conclude that

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{k^2} + (\frac{1}{3}k + 2.4)\delta$$

and we retrieve the classical behavior of a fast-gradient method when used with a $(\delta, L) = (\delta, L, 0)$-oracle (see [24]):

- ◇ A needed number of iterations in $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$
- ◇ A needed oracle accuracy of order at least $\delta = \Theta\left(\frac{\epsilon^{3/2}}{L^{1/2}R}\right)$.

Now we can obtain the following convergence rate for the fast gradient method using a (δ, L, μ)-oracle

Theorem 5.14. *The fast gradient method applied to a function f endowed with a (δ, L, μ)-oracle generates a sequence $\{y_k\}_{k\geq 1}$ satisfying:*

$$\begin{aligned} f(y_k) - f^* \quad &\leq \quad \min\left(\frac{4Ld(x^*)}{k^2}, Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\quad + \min\left((\frac{1}{3}k + 2.4), \left(1 + \sqrt{\frac{L}{\mu}}\right)\right)\delta. \end{aligned}$$

Proof. Using the Lemmas 5.10 and 5.12 and the remarks 5.11 and 5.12 in the convergence rate given by the Theorem 5.9, we obtain

$$f(y_k) - f^* \leq \min\left(\frac{4Ld(x^*)}{k^2}, \frac{Ld(x^*)}{(1+\frac{1}{2}\sqrt{\frac{\mu}{L}})^{2k}}\right) + \min\left((\frac{1}{3}k + 2.4), \left(1 + \sqrt{\frac{L}{\mu}}\right)\right)\delta.$$

As for $x \in [0, \frac{1}{4}]$, we have that $\log(1+2x) \geq x$:

$$\begin{aligned} \frac{1}{(1+\frac{1}{2}\sqrt{\frac{\mu}{L}})^{2k}} \quad &= \quad \exp\left(-2k\log\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\leq \quad \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right). \end{aligned}$$

We conclude that

$$\begin{aligned} f(y_k) - f^* &\leq \min\left(\frac{4Ld(x^*)}{k^2}, Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\quad + \min\left((\frac{1}{3}k + 2.4), \left(1 + \sqrt{\frac{L}{\mu}}\right)\right)\delta. \end{aligned}$$

□

5.5.3 Oracle accuracy fixed: Best reachable target accuracy using the FGM

We have seen that, contrarily to the GM, the FGM suffers from a problem of error increase. The extra error in the convergence rate, due to the (δ, L, μ) oracle, is not δ but something of order $\min\left(k\delta, \sqrt{\frac{L}{\mu}}\delta\right)$. As a consequence, the best possible level of accuracy δ cannot be reached by the FGM. In this section, we are interested in the best accuracy ϵ that we can obtain for $f(y_k) - f^*$, using the FGM with a (δ, L, μ) oracle.

We ensure here that δ, L, μ and R are fixed quantities. The only degree of freedom that we have is the number of iterations k that we perform. In Theorem 5.14, we have obtained the following model for the convergence rate of the FGM when applied to a function endowed with a (δ, L, μ) oracle:

$$f(y_k) - f^* \leq F(k) := \min\left(F_1(k), F_2(k), F_3(k), F_4(k)\right)$$

where

1. $F_1(k) = \frac{4Ld(x^*)}{k^2} + \left(\frac{1}{3}k + 2.4\right)\delta$
2. $F_2(k) = \frac{4Ld(x^*)}{k^2} + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$
3. $F_3(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(\frac{1}{3}k + 2.4\right)\delta$
4. $F_4(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta.$

The minimum of $F_1(k)$ is reached at $k_1^* = \Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$ and $F_1^* = F_1(k_1^*) = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$.

The minimum of $F_2(k)$, $F_2^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$, is reached at the limit. However we can obtain accuracy of the same order after $k_2^* = \Theta\left(\frac{(L\mu)^{1/4}R}{\delta^{1/2}}\right)$ iterations.

The minimum of $F_3(k)$ is reached at $k_3^* = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$ and $F_3^* = F_3(k_3^*) = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\left(\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)+1\right)\right)$.
The minimum of $F_4(k)$, $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ is reached at the limit. However we can obtain accuracy of the same order after $k_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$ iterations.

We conclude that the best reachable accuracy by the FGM is of order $\min\{F_1^*, F_4^*\} = \min\{\Theta(L^{1/3}R^{2/3}\delta^{2/3}), \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)\}$. More precisely, we can consider two different situations:

- ◇ A) The condition number $Q = \frac{L}{\mu}$ is sufficiently small and/or the oracle accuracy δ is sufficiently small and/or R is sufficiently big such that $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right) \leq F_1^* = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ i.e. $\delta \leq \Theta\left(\frac{R^2\mu^{3/2}}{L^{1/2}}\right)$.
  In this situation, we can gain from the fact that $\mu > 0$ and reach a level of accuracy $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ better than the best level reachable in the case $\mu = 0$ (i.e. $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$). The number of iterations needed in order to reach this accuracy is of order $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$.

  Remark 5.13. We also can reach the level $\Theta(L^{1/3}R^{2/3}\delta^{2/3})$ and the needed number of iterations is $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)\right)$. This number of iterations is smaller or of the same order as $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$, the needed number of iterations by the FGM to reach this accuracy in the case $\mu = 0$.

  Remark 5.14. We can also reach the level of accuracy $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ using the GM. But the needed number of iteration is of order $\Theta\left(\frac{L}{\mu}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$, worse than $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$.

- ◇ B) The condition number $Q = \frac{L}{\mu}$ is sufficiently big and/or the oracle accuracy δ is sufficiently big and/or R is sufficiently small such that $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right) \geq F_1^* = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ i.e. $\delta \geq \Theta\left(\frac{R^2\mu^{3/2}}{L^{1/2}}\right)$.
  In this situation, we cannot exploit the fact that $\mu > 0$. The best reachable accuracy is of level $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ after $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$ iterations. It is the same result as what we had obtained in Chapter 4 in the case $\mu = 0$.

Remark 5.15. We can reach also reach the level $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$. The needed number of iteration is $\Theta\left(\frac{(L\mu)^{1/4}R}{\delta^{1/2}}\right)$.

Remark 5.16. We can also reach the level of accuracy $F_1^* = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$ using the GM. But the needed number of iterations is of order $\min\{\Theta\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)\right)\} = \Theta\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)$ which is worse than $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$.

Remark 5.17. The impossibility to exploit the fact that $\mu > 0$ when μ is too small comes perhaps from our analysis. More precisely, it might come from the fact that we have bounded $f(y_k) - f^* \leq \frac{1}{A_k}\left(d(x^*) + \sum_{i=0}^{k} A_i\delta\right)$ using the two approximations:

1. $1 + \mu A_k \approx 1$ when μ is small
2. $1 + \mu A_k \approx \mu A_k$ when μ is big.

It would seems more natural for the best reachable accuracy, to be a continuous function with limits $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ when $\mu \to +\infty$ and $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ when $\mu \to 0$. However, it seems very difficult to find an upper bound for $f(y_k) - f^*$ which is at the same time accurate and easy to analyze.

In conclusion, the best accuracy reachable by the FGM when endowed with a (δ, L, μ) oracle is of order $\min\{\Theta(L^{1/3}R^{2/3}\delta^{2/3}), \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)\}$. If such accuracy is sufficient, it is preferable to use the FGM instead of the GM. However, if we want to reach a better level of accuracy, for example of order $\Theta(\delta)$, the only possibility is to use the GM, slower but less sensitive to the oracle error.

5.5.4 Oracle accuracy not fixed: Required number of iterations and required oracle accuracy for a given target accuracy

In this subsection, we consider a different situation. We assume that we can choose the number of iterations k and the oracle accuracy δ but that we want to reach a target accuracy ϵ for the objective function. Furthermore, in this case we assume that L and μ are independent of the oracle accuracy δ. A way to ensure $f(y_k) - f^* \leq \epsilon$ is to choose k and δ such that one of the four models $F_i(k)$ is smaller than ϵ. First, we will consider the four models separately. Each model contains three terms and we will ensure $F_i(k) \leq \epsilon$ by imposing that each term in the model is smaller than $\frac{\epsilon}{3}$ (another repartition of the desired accuracy

between the different terms of a model leads simply to different constant factors in the resulting expressions for k and δ).

1. Model 1: $F_1(k) = \frac{4Ld(x^*)}{k^2} + \left(\frac{1}{3}k + 2.4\right)\delta$.
   We have the three conditions ensuring $F_1(k) \leq \epsilon$:
   - $$\frac{4Ld(x^*)}{k^2} \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}.$$
     We choose $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$
   - $$\frac{1}{3}k\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}}$$
   - $$2.4\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon}{7.2}.$$

   Therefore a first possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$ iterations with $\delta = \min\{\frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}}, \frac{\epsilon}{7.2}\}$.

2. Model 2: $F_2(k) = \frac{4Ld(x^*)}{k^2} + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$.
   We have the three conditions ensuring $F_2(k) \leq \epsilon$:
   - $$\frac{4Ld(x^*)}{k^2} \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
   - $$\delta \leq \frac{\epsilon}{3}$$
   - $$\sqrt{\frac{L}{\mu}}\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon.$$

   Therefore a second possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$ iterations with $\delta = \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon$.

3. Model 3: $F_3(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(\frac{1}{3}k + 2.4\right)\delta$.
   We have the three conditions ensuring $F_3(k) \leq \epsilon$:
   - $$Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right).$$
     We choose $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$.

◇
$$\frac{1}{3}k\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$

◇
$$2.4\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon}{7.2}.$$

Therefore a third condition ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$ iterations with $\delta = \min\{\frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}, \frac{\epsilon}{7.2}\}$.

4. Model 4: $F_4(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$. We have the three conditions ensuring $F_4(k) \leq \epsilon$:

◇
$$Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$$

◇
$$\delta \leq \frac{\epsilon}{3}$$

◇
$$\sqrt{\frac{L}{\mu}}\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon.$$

Therefore a fourth possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$ with $\delta = \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon$.

Of course, we want to reach $f(y_k) - f^* \leq \epsilon$ in a minimum number of iterations k and with δ as big as possible (δ representing the accuracy of the first-order information, it seems natural that a high accuracy for δ is costly). We will choose between these four possibilities that ensure $f(y_k) - f^* \leq \epsilon$ with the minimization of k as a first criterion and the maximization of δ as a second criterion.

Remark 5.18. For simplicity, we assume here that $\epsilon \leq 0.2315Ld(x^*)$ i.e. $\frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}} \leq \frac{\epsilon}{7.2}$. In particular, this assumption implies:

$$\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \geq \frac{3}{2}.$$

We consider two main cases :

1. **Case 1:**
$$2\sqrt{\frac{3Ld(x^*)}{\epsilon}} \leq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$$
i.e.
$$\sqrt{\frac{L}{\mu}} \geq \frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$
In this case, models 1 and 2 are the most favorable regarding to the number of iterations. We perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}} = \Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations. Concerning the needed oracle accuracy, we have to consider two different subcases:
   - ◇ **Case 1.1:**
$$\frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)} \leq \sqrt{\frac{L}{\mu}} \leq \frac{2}{3}\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
In this case $\sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} \geq \frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}}$ and model 2 is more interesting than the first one. We choose $\delta = \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$.
   - ◇ **Case 1.2:**
$$\sqrt{\frac{L}{\mu}} \geq \frac{2}{3}\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
In this case $\frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}} \geq \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3}$ and model 1 is more interesting than the second one. We choose $\delta = \frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}} = \Theta\left(\frac{\epsilon^{3/2}}{L^{1/2}R}\right)$.

Remark 5.19. : In the case 1.2., we do not exploit the fact that $\mu > 0$. We obtain the same number of iterations and the same oracle accuracy in the case $\mu = 0$.

2. **Case 2:**
$$2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \leq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
i.e.:
$$\sqrt{\frac{L}{\mu}} \leq \frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$

In this case, models 3 and 4 are the most favorable with respect to the needed number of iterations. We perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right) = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{3LR^2}{\epsilon}\right)\right) iterations$. As $\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \geq \frac{3}{2}$, we have

$$\frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)} \leq \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3}.$$

Therefore model 4 is always more interesting than the third one and we choose $\delta = \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$.

5.6 GM and FGM for uniformly convex problems with different levels of smoothness

In the Sections 5.3 and 5.5, we have studied the effect of a (δ, L, μ) oracle on the GM and the FGM. In particular, we have established the complexity of these methods in a inexact framework and the link between desired final accuracy and needed oracle accuracy.

In subsection 5.2.5, we have seen that a (δ, L, μ) oracle can be also available for functions that are not in $S^{1,1}_{\mu,L}(Q)$. More precisely, the function can be uniformly convex (instead of strongly convex) and nonsmooth or weakly smooth (instead of smooth). The exact oracle for a function $f \in U^{0,\rho,\nu}_{\kappa,M}(Q)$ can be seen as a (δ, L, μ) oracle.

If we put these results together, we can apply the GM and the FGM, initially designed for functions in $S^{1,1}_{\mu,L}(Q)$, to functions in $U^{0,\rho,\nu}_{\kappa,M}(Q)$. In this section, we study the complexity of these two methods on various classes of convex problems with different levels of smoothness and different levels of uniform convexity. For simplicity, we are only interested in the order of the dependence of these complexities on ϵ (the desired final accuracy), κ and M, not on the absolute constant factors. Furthermore, Theorem 5.3 is applied with $\delta_1 = \delta_2 = \frac{\delta}{2}$.

5.6.1 Gradient Method for function in $U^{0,\rho,\nu}_{\kappa,M}(Q)$.

If we apply the gradient method to a function endowed with a (δ, L, μ) oracle and if the desired accuracy is $\epsilon > 0$, we know that the number of iterations that we have to perform is

$$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)$$

with an oracle accuracy $\delta = \Theta(\epsilon)$. When the (δ, L, μ) oracle is in fact an exact oracle of a function $f \in U^{0,\rho,\mu}_{\kappa,M}(Q)$, we have (see Theorem 5.3) $L = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\delta^{\frac{1-\nu}{1+\nu}}}\right)$ and $\mu = \Theta\left(\kappa^{\frac{2}{\rho}}\delta^{\frac{\rho-2}{\rho}}\right)$. Therefore

$$\frac{L}{\mu} = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\delta^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\right)$$

and

$$\log\left(\frac{LR^2}{\epsilon}\right) = \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

We obtain the complexity:

$$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

We can particularize this complexity bound for different classes of uniformly convex problems:

1. The smooth case $\nu = 1$ (f has a Lipschitz-continuous gradient)
   - ◇ Strong convexity $\rho = 2$:
     $$\Theta\left(\frac{M}{\kappa}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$
     We retrieve the non-optimal complexity of the gradient method on $S^{1,1}_{\kappa,M}(Q)$.
   - ◇ Uniform convexity $\rho \geq 2$:
     $$\Theta\left(\frac{M}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{\rho-2}{\rho}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$
     This complexity cannot be optimal (see what we obtain in the next subsection with the FGM).
2. The nonsmooth case $\nu = 0$ (f has subgradients with bounded variation)
   - ◇ Strong convexity $\rho = 2$:
     $$\Theta\left(\frac{M^2}{\kappa\epsilon}\log\left(\frac{M^2R^2}{\epsilon^2}\right)\right)$$
     This complexity is optimal up to a logarithmic factor (see [55, 33]).

◇ Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{M^2}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{2(\rho-1)}{\rho}}}\log\left(\frac{M^2R^2}{\epsilon^2}\right)\right)$$

This complexity is optimal, up to a logarithmic factor (see [33]).

3. The weakly smooth case $0 < \nu < 1$ (f has a Hölder continuous gradient)

◇ Strong convexity $\rho = 2$:

$$\Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa\epsilon^{\frac{1-\nu}{1+\nu}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right)$$

This complexity cannot be optimal (see what we obtain with FGM in the next subsection).

◇ Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

5.6.2 Fast Gradient Method for functions in $U^{0,\rho,\nu}_{\kappa,M}(Q)$

If we apply the fast gradient method to a function endowed with a (δ, L, μ) oracle and if the desired accuracy is ϵ, we know that the number of iterations that we have to perform is proportional to

$$\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)$$

with an oracle accuracy $\delta = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$. When the (δ, L, μ) oracle is in fact an exact oracle of a function $f \in U^{0,\rho,\mu}_{\kappa,M}(Q)$, we have (see Theorem 5.3) $L = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\delta^{\frac{1-\nu}{1+\nu}}}\right)$ and $\mu = \Theta\left(\kappa^{\frac{2}{\rho}}\delta^{\frac{\rho-2}{\rho}}\right)$. We obtain:

$$\sqrt{\frac{L}{\mu}} = \Theta\left(\frac{M^{\frac{1}{1+\nu}}}{\kappa^{\frac{1}{\rho}}\delta^{\frac{1}{2}\left(\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}\right)}}\right)$$

and therefore

$$\sqrt{\frac{L}{\mu}}\delta = \Theta\left(\frac{M^{\frac{1}{1+\nu}}\delta^{1-\frac{1}{2}\left(\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}\right)}}{\kappa^{\frac{1}{\rho}}}\right).$$

As $\sqrt{\frac{L}{\mu}}\delta = \Theta(\epsilon)$ and as $1 - \frac{1}{2}\left(\frac{1-\nu}{1+\nu} + \frac{\rho-2}{\rho}\right) = \frac{\nu}{\nu+1} + \frac{1}{\rho}$, we obtain that $\delta = \Theta\left(\left(\frac{\kappa^{\frac{1}{\rho}}\epsilon}{M^{\frac{1}{1+\nu}}}\right)^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\right)$. Therefore, we have

$$\sqrt{\frac{L}{\mu}} = \Theta\left(\frac{\epsilon}{\delta}\right) = \Theta\left(\frac{M^{\frac{1}{1+\nu}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}}{\kappa^{\frac{1}{\rho}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\epsilon^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}-1}}\right)$$

and

$$\begin{aligned}\log\left(\frac{LR^2}{\epsilon}\right) &= \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}}}{\left(\left(\frac{\kappa^{\frac{1}{\rho}}\epsilon}{M^{\frac{1}{1+\nu}}}\right)^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\right)^{\frac{1-\nu}{1+\nu}}}\frac{R^2}{\epsilon}\right)\right)\\ &= \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}+\frac{1-\nu}{(1+\nu)^2}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}R^2}{\kappa^{\frac{1}{\rho}\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}\epsilon^{\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)+1}}\right)\right).\end{aligned}$$

We obtain the complexity:

$$\Theta\left(\left(\frac{M^{\rho}}{\kappa^{\nu+1}\epsilon^{\rho-\nu-1}}\right)^{\frac{1}{\nu(\rho+1)+1}}\log\left(\left(\frac{M^{2+\rho}}{\kappa^{1-\nu}\epsilon^{\rho+\nu+1}}\right)^{\frac{1}{\nu(\rho+1)+1}}R^2\right)\right).$$

We particularize now this complexity bound on different classes of uniformly convex problems:

1. The smooth case $\nu = 1$ (f has a Lipschitz-continuous gradient)
   - ◇ Strong convexity $\rho = 2$:

     $$\Theta\left(\frac{M^{\frac{1}{2}}}{\kappa^{\frac{1}{2}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$

     We retrieve the optimal complexity of the fast gradient method on $S^{1,1}_{\kappa,M}(Q)$.
   - ◇ Uniform convexity $\rho \geq 2$:

     $$\Theta\left(\frac{M^{\frac{\rho}{\rho+2}}}{\kappa^{\frac{2}{\rho+2}}\epsilon^{\frac{\rho-2}{\rho+2}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$

2. The nonsmooth case $\nu = 0$ (f has subgradients with bounded variation)

   ◇ Strong convexity $\rho = 2$:

   $$\Theta\left(\frac{M^2}{\kappa\epsilon}\log\left(\frac{M^4R^2}{\kappa\epsilon^3}\right)\right)$$

   This complexity is optimal up to a logarithmic factor (see [55, 33]).

   ◇ Uniform convexity $\rho \geq 2$:

   $$\Theta\left(\frac{M^\rho}{\kappa\epsilon^{\rho-1}}\log\left(\frac{M^{2+\rho}R^2}{\kappa\epsilon^{\rho+1}}\right)\right)$$

   This complexity is clearly non-optimal (compare with what we obtain using the GM in the previous subsection).

3. The weakly smooth case $0 < \nu < 1$ (f has a Hölder continuous gradient)

   ◇ Strong convexity $\rho = 2$:

   $$\Theta\left(\frac{M^{\frac{2}{1+3\nu}}}{\kappa^{\frac{\nu+1}{1+3\nu}}\epsilon^{\frac{1-\nu}{1+3\nu}}}\log\left(\frac{M^{\frac{4}{(1+3\nu)}}R^2}{\kappa^{\frac{1-\nu}{1+3\nu}}\epsilon^{\frac{3+\nu}{1+3\nu}}}\right)\right)$$

   ◇ Uniform convexity $\rho \geq 2$:

   $$\Theta\left(\left(\frac{M^\rho}{\kappa^{\nu+1}\epsilon^{\rho-\nu-1}}\right)^{\frac{1}{\nu(\rho+1)+1}}\log\left(\left(\frac{M^{2+\rho}}{\kappa^{1-\nu}\epsilon^{\rho+\nu+1}}\right)^{\frac{1}{\nu(\rho+1)+1}}R^2\right)\right).$$

5.7 Lower bound on error increase

Applicability of first-order methods, initially designed for smooth strongly convex problems, to nonsmooth strongly convex problems, using the notion of inexact oracle, opens a possibility to derive lower bounds on error increase. This is the main subject of this section. We start from the following observation.

Theorem 5.15. *Consider a first-order method for $S^{1,1}_{\mu,L}(Q)$. Assume that the bounds on the performance of this method, as applied to a problem equipped with a (δ, L, μ)-oracle, are given by inequality*

$$\begin{aligned} f(z_k) - f^* \quad \leq \quad & \min\left(C_1\tfrac{LR^2}{k^{p_1}}, C_2LR^2\exp\left(-k\left(\tfrac{\mu}{L}\right)^{p_2}\right)\right) \\ & + \min\left(C_3k^{q_1}\delta, C_4\left(\tfrac{L}{\mu}\right)^{q_2}\delta\right). \end{aligned} \tag{5.18}$$

where C_1, C_2, C_3, C_4 are absolute constants, and k is the iteration counter. Then the inequalities

$$q_1 \geq p_1 - 1$$

and

$$q_2 \geq 1 - p_2$$

must hold.

Proof. $\diamond$ $q_1 \geq p_1 - 1$.

Let f be a nonsmooth convex function, whose subgradients have variation bounded by constant M i.e $f \in F_M^{0,0}(Q)$. We have seen in Chapter 4 that for such a function, the standard oracle can be treated as a $(\delta, \frac{M^2}{2\delta}, 0)$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$f(z_k) - f^* \quad \leq \quad \frac{C_1 M^2 R^2}{2\delta k_1^p} + C_3 k^{q_1} \delta.$$

Optimizing the right-hand side of this inequality in δ, we get

$$f(z_k) - f^* \quad \leq \quad [2C_1C_3]^{1/2} MR \cdot k^{-\frac{p_1 - q_1}{2}}.$$

From the lower complexity bounds for nonsmooth optimization problems, we know that black-box methods cannot converge faster than $O(\frac{1}{k^{1/2}})$. Hence, we conclude that $p_1 - q_1 \leq 1$.

$\diamond$ $q_2 \geq 1 - p_2$.

Let f be a nonsmooth strongly convex function, whose subgradients have variation bounded by constant M i.e. $f \in U_{\mu,M}^{1,2,0}(Q)$. We have seen in Theorem 5.3 that for such a function, the standard oracle can be treated as $(\delta, \frac{M^2}{2\delta}, \mu)$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$\begin{aligned} f(z_k) - f^* \quad &\leq \quad \frac{C_2 M^2 R^2}{2\delta} \exp\left(-k\left(\frac{\mu}{M^2} 2\delta\right)^{p_2}\right) + C_4 \left(\frac{M^2}{2\delta\mu}\right)^{q_2} \delta \\ &= \quad \frac{C_2 M^2 R^2}{2\delta} \exp\left(-k \frac{\mu^{p_2} 2^{p_2} \delta^{p_2}}{M^{2p_2}}\right) + C_4 \frac{M^{2q_2}}{2^{q_2} \mu^q} \delta^{1-q_2}. \end{aligned}$$

If we choose δ such that $C_4 \frac{M^{2q_2}}{2^{q_2}\mu^{q_2}} \delta^{1-q_2} = \frac{\epsilon}{2}$, we obtain $\delta(\epsilon) = \frac{1}{2} \frac{\mu^{\frac{q_2}{1-q_2}} \epsilon^{\frac{1}{1-q_2}}}{C_4^{\frac{1}{1-q_2}} M^{\frac{2q_2}{1-q_2}}}$. Therefore, if we want an accuracy of ϵ for the objective function, we can choose k such that $\frac{C_2 M^2 R^2}{2\delta(\epsilon)} \exp\left(-k \frac{\mu^{p_2} 2^{p_2} \delta^{p_2}}{M^{2p_2}}\right) = \frac{\epsilon}{2}$ i.e.

$$k = \frac{M^{\frac{2p_2}{1-q_2}} C_4^{\frac{p_2}{1-q_2}}}{\mu^{\frac{p_2}{1-q_2}} \epsilon^{\frac{p_2}{1-q_2}}} \log\left(\frac{2C_2 M^{\frac{2}{1-q_2}} R^2}{\epsilon^{\frac{2-q_2}{1-q_2}} \mu^{\frac{q_2}{1-q_2}}}\right).$$

From the lower complexity bounds for nonsmooth strongly convex optimization problems, we know that black-box methods cannot have a better complexity that $O\left(\frac{1}{\epsilon}\right)$ (see [55, 33]). Hence, we conclude that

$$\frac{p_2}{1-q_2} \geq 1 \quad \Leftrightarrow p_2 \geq 1 - q_2.$$

□

We can consider two extreme cases:

- $\diamond$ $q_1 = 0$ and $q_2 = 0 \Rightarrow p_1 \leq 1$ and $p_2 \geq 1$:
  It is impossible to have a first-order method without increase of errors, which has better complexity than GM, that is
  $\min\left(\Theta\left(\frac{LR^2}{\epsilon}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$.

- $\diamond$ $p_1 = 2$ and $p_2 = \frac{1}{2} \Rightarrow q_1 \geq 1$ and $q_2 \geq \frac{1}{2}$:
  If we want a first-order method with optimal complexity
  $\min\left(\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right), \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$ like the FGM, then it must suffer from increase of errors, with factor at least of order
  $\min\left(\Theta(k), \Theta\left(\sqrt{\frac{L}{\mu}}\right)\right)$ (we obtain exactly this factor for the FGM).

Chapter 6

Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle

This chapter corresponds to the paper [27]:
O. Devolder, F. Glineur and Yu. Nesterov. **Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle.** *CORE Discussion Paper 2013/17*.

Chapter 6 in two Questions/ Answers

◇ *Is there something missing between the robust but slow PGM/DGM and the fast but sensitive FGM ?*

Yes, clearly! When used with inexact information, the FGM cannot reach good target accuracies which often obliges us to come back to the slow PGM or DGM. We cannot be satisfied with this situation. Of course, we know that fastness and sensitivity to errors are linked. But, between the two extreme choices of the robust but slow PGM/DGM and the fast but sensitive FGM, it could be preferable to find a good compromise, a method with intermediate speed and intermediate sensitivity to errors.

◇ *Is it possible to develop new methods having a behavior optimized in view of the level of errors in the first-order information ?*

Yes. We have developed a new family of methods, the Intermediate Gradient Methods (IGM), that are able to exhibit any intermediate behavior between the PGM/DGM and the FGM. In particular, for a given target accuracy ϵ and a given oracle error δ, we can optimize the behavior of

the method in order to obtain an optimal trade-off between speed of convergence and robustness. That allows us to reach accuracies unreachable by the FGM in a significantly smaller amount of iterations compared to what is needed with the PGM/DGM.

Contents

6.1 The Intermediate Gradient Method (IGM) **189**
6.1.1 General Intermediate Gradient Method 189
6.1.2 Variant with prox-type subproblems 194
6.2 Link with existing methods **197**
6.2.1 Link with Fast Gradient Method 197
6.2.2 Link with Dual Gradient Method 198
6.3 Optimal choice of the coefficients **199**
6.3.1 Optimal Policy 199
6.3.2 Lower Complexity bound 203
6.4 Switching policy for the coefficients **204**
6.4.1 Feasible Policy 205
6.4.2 Optimal Switching Level l 207
6.4.3 Optimal Switching Moment m 208
6.4.4 General optimality of the optimal switching policy 215
6.4.5 Conclusion: Optimal Switching Policy 217
6.5 Improvement compared with existing methods **218**
6.6 Power policy for the coefficients **227**
6.6.1 Power Policy 227
6.6.2 Optimal power choice 229
6.7 Numerical Illustration **232**
6.7.1 Behavior with exact oracle $\delta = 0$ 233
6.7.2 Introducing errors in the first-order methods: behavior with $\delta = 1e-2$ 234
6.7.3 Increasing the oracle errors: behavior with $\delta = 1e-1$ 236

The analysis of existing first-order methods of smooth convex optimization when used with inexact information shows us (see Chapter 4) that:

- ◇ The GM can be seen as a slow method but robust with respect to oracle errors. Used with a (δ, L)-oracle, the convergence rate of the GM become:

$$f(y_k) - f^* \leq O\left(\frac{LR^2}{k}\right) + \delta.$$

The method is slow but there is no accumulation of errors. The upper-bound for the objective function decreases with k and asymptotically tends to δ.

Any target accuracy ϵ over δ can be reached by the GM and the corresponding needed number of iteration is given by $\frac{LR^2}{\epsilon-\delta}$.

- ◇ The FGM can be seen as a fast method but sensitive with respect to oracle errors. Indeed, when used with a (δ, L) oracle, the FGM exhibits convergence rate

$$f(y_k) - f^* \leq O\left(\frac{LR^2}{k^2}\right) + O(k\delta).$$

Contrarily to the GM, the use of inexact oracle in FGM results in error accumulation. Indeed, while the first term decreases as $O(\frac{1}{k^2})$, the second term is increasing with k and the FGM, when used with an inexact oracle, is asymptotically divergent.

The error on $f(y_k) - f^*$ attains its minimum value after $\Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ iterations with corresponding best reachable accuracy

$$\epsilon^*_{FGM} = \Theta(L^{1/3}R^{2/3}\delta^{2/3}) > \delta.$$

It means that when looking for a relatively less accurate solution $\epsilon > \epsilon^*_{FGM}$, we can still use the FGM and the corresponding needed number of iterations $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ is much more reasonable that what is needed by the GM.

But if we want to obtain an accuracy below the threshold ϵ^*_{FGM} (i.e. $\delta \leq \epsilon < \epsilon^*_{FGM}$), we cannot use the FGM anymore and we need to come back to the slow GM.

At this step, it seems clear that we cannot stay with this pessimistic observation. There is something missing between the GM and the FGM. We need to develop new first-order methods that, when used with a (δ, L)-oracle, are faster than the GM but that can reach accuracy below ϵ^*_{FGM}.

Ideally, we would like to obtain a method which shares the best of the GM and of the FGM, a method which is as fast as the FGM (i.e. with a convergence rate

$\Theta\left(\frac{LR^2}{k^2}\right)$ in the exact case) and as robust with respect to the oracle error as the GM (i.e. without accumulation of error). In term of complexity, it would mean to obtain a method that can reach any accuracy over δ in only $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations.

Unfortunately, this goal is too ambitious. The fastness of a first-order method and its sensitivity with respect to errors are linked. Theorem 4.12 of Chapter 4 shows us that accumulation of errors is an intrinsic and unavoidable property of any fast first-order method using inexact oracle.

This result is not good news: there is no hope to develop a first-order method which is at the same time as fast as the FGM and as robust as the GM. There is no free lunch: the faster the method is, the higher its sensitivity to error is.

However, this is not the end of the story. Between the two extreme choices of the robust but slow GM and the fast but highly sensitive FGM, it could be preferable to develop methods with intermediate speed and intermediate sensitivity to errors. This is the goal of this chapter: we want to develop methods with intermediate behaviors between GM and FGM methods.

The structure of this chapter looks as follows:

In the following section, we develop a general family of first-order methods, the IGM (Intermediate Gradient Method) which is mainly based on two sequences of coefficients $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$. The degrees of freedom in the choice of these coefficients allows us to obtain different intermediate behaviors.

In Section 6.2, we see that the existing DGM and FGM are nothing else that an IGM but with particular choices of the coefficients. In Section 6.3, given an oracle accuracy δ, we are interested in the optimal coefficient choices for reaching a target accuracy $\epsilon > 0$, i.e. the choice of coefficients that allows us to reach the accuracy ϵ with a minimum number of iterations. We derive important properties that such optimal coefficient policy must satisfy and derive lower bound on the complexity that we can expect with an optimal choice of the coefficients.

In Section 6.4 and 6.5, we propose a practical coefficient policy that matches our lower bound and allows us to reach a target accuracy $\epsilon < \epsilon^*_{FGM}$ with a (significantly) smaller number of iterations compared to what is needed using the GM. In Section 6.6, with another choice for the coefficients, we are able to generate methods exhibiting the whole spectrum of convergence rates given by Theorem 4.12, corresponding to every possible trade-off between fastness

of the method and robustness to errors. In the last section (Section 6.7), we present a numerical illustration of the results obtained in this chapter.

6.1 The Intermediate Gradient Method (IGM)

6.1.1 General Intermediate Gradient Method

Let $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying for all $i \geq 0$

$$\alpha_i^2 \leq B_i \leq \sum_{j=0}^{i} \alpha_j \tag{6.1}$$

$$0 \leq \alpha_i \leq B_i. \tag{6.2}$$

We define also $A_i = \sum_{j=0}^{i} \alpha_j$ and $\tau_i = \frac{\alpha_{i+1}}{B_{i+1}}$.

For a given choice of the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ and of the prox-function d, the corresponding Intermediate Gradient Method (IGM) looks as follows when applied to an objective function f endowed with a (δ, L)-oracle:

Algorithm 19 Intermediate Gradient Method (IGM)

1: Compute $x_0 = \arg\min_{x\in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$
4: Compute

$$w_k = \arg\min_{x\in Q}\{\frac{L}{2}\|x - x_k\|_E^2 + \langle g_{\delta,L}(x_k), x - x_k\rangle\} \tag{6.3}$$

5: Compute

$$z_k = \arg\min_{x\in Q}\{Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i\rangle\} \tag{6.4}$$

6: Compute

$$y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k \tag{6.5}$$

7: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
8: **end for**

Remark 6.1. We have $B_0 = A_0 = \alpha_0$ and therefore $y_0 = w_0$. We do not need to define y_{-1}.

Remark 6.2. Due to the condition $0 \leq B_k \leq A_k$, we know that $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$ is a convex combination. In the same way, the condition $0 \leq \alpha_k \leq B_k$ ensures that $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$ is also a convex combination.

Let us establish now the convergence rate of this Intermediate Gradient Method. We start first with the following lemma:

Lemma 6.1. *Assume that the IGM is applied to a function f endowed with a (δ, L)-oracle. Then for all $k \geq 0$, we have*

$$A_k f(y_k) \leq \Psi_k^* + E_k$$

where $E_k = \left(\sum_{i=0}^{k} B_i \right) \delta$ *and*

$$\Psi_k^* = \min_{x \in Q} \{ \Psi_k(x) := L d(x) + \sum_{i=0}^{k} \alpha_i [f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i \rangle] \}.$$

Proof. Denote $f_k = f_{\delta,L}(x_k), \quad g_k = g_{\delta,L}(x_k)$.

- ◇ **For $k = 0$:**
  Since $\alpha_0 \leq 1$, we have

$$\begin{aligned} \Psi_0^* &= \min_{x \in Q} \{ L d(x) + \alpha_0 [f_0 + \langle g_0, x - x_0 \rangle] \} \\ &\overset{(2.8)}{\geq} \alpha_0 \min_{x \in Q} \{ f_0 + \langle g_0, x - x_0 \rangle + \frac{L}{2} \|x - x_0\|_E^2 \} \\ &\overset{(4.2)}{\geq} \alpha_0 f(y_0) - \alpha_0 \delta. \end{aligned}$$

  Therefore, we have:

$$\alpha_0 f(y_0) \leq \Psi_0^* + \alpha_0 \delta = \Psi_0^* + B_0 \delta.$$

- ◇ **If it is true for $k \geq 0$:**
  By the optimality condition of the optimization problem defining z_k:

$$\langle L \nabla d(z_k) + \sum_{i=0}^{k} \alpha_i g_i, x - z_k \rangle \geq 0.$$

Hence in view of strong convexity of d:

$$\begin{aligned} d(x) &\geq d(z_k) + \langle \nabla d(z_k), x - z_k \rangle + \frac{1}{2} \|x - z_k\|_E^2 \\ &\geq d(z_k) + \sum_{i=0}^{k} \frac{\alpha_i}{L} \langle g_i, z_k - x \rangle + \frac{1}{2} \|x - z_k\|_E^2 . \end{aligned}$$

Thus for all $x \in Q$:

$$\begin{aligned} & Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i \rangle] \\ \geq\ & Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i \rangle] \\ & + \frac{L}{2} \|x - z_k\|_E^2 + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \end{aligned}$$

We obtain

$$\Psi_{k+1}^* \overset{(6.4)}{\geq} \Psi_k^* + \min_{x \in Q} \{ \frac{L}{2} \|x - z_k\|_E^2 + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \}. \tag{6.6}$$

On the other hand, as $A_k = (B_{k+1} - \alpha_{k+1}) + (A_k - B_{k+1} + \alpha_{k+1}) = (B_{k+1} - \alpha_{k+1}) + (A_{k+1} - B_{k+1})$ and as we assume that $\Psi_k^* \geq A_k f(y_k) - E_k$, we have

$$\begin{aligned} & \Psi_k^* + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ \geq\ & A_k f(y_k) - E_k + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ =\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + (B_{k+1} - \alpha_{k+1}) f(y_k) \\ & + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ \overset{(4.2)}{\geq}\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + (B_{k+1} \\ & - \alpha_{k+1}) [f_{k+1} + \langle g_{k+1}, y_k - x_{k+1} \rangle] \\ & + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ =\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} + \\ & \langle g_{k+1}, (B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1}) \rangle. \end{aligned}$$

As $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$ and $\tau_k = \frac{\alpha_{k+1}}{B_{k+1}}$, we have also

$$(B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1}) = \alpha_{k+1}(x - z_k).$$

We obtain that

$$\begin{aligned}&\Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ &\quad \geq (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle.\end{aligned} \tag{6.7}$$

Therefore using the equations (6.6) and (6.7), we have

$$\begin{aligned}\Psi_{k+1}^* &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} \\ &\quad + \min_{x\in Q}\{\frac{L}{2}\|x - z_k\|_E^2 + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle\} \\ &= (A_{k+1} - B_{k+1})f(y_k) - E_k \\ &\quad + B_{k+1}[f_{k+1} + \min_{x\in Q}\{\frac{L}{2B_{k+1}}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}].\end{aligned}$$

As $\alpha_{k+1}^2 \leq B_{k+1}$, we have $\tau_k^2 \leq \frac{1}{B_{k+1}}$ and we conclude that

$$\begin{aligned}\Psi_{k+1}^* &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}[f_{k+1} \\ &\quad + \min_{x\in Q}\{\frac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\}].\end{aligned}$$

For $x \in Q$, let us now define

$$y = \tau_k x + (1 - \tau_k)y_k.$$

Since $y - x_{k+1} = \tau_k(x - z_k)$, we obtain

$$\begin{aligned}&\min_{x\in Q}\{\frac{\tau_k^2 L}{2}\|x - z_k\|_E^2 + \tau_k\langle g_{k+1}, x - z_k\rangle\} \\ &= \min_y\{\frac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle : y \in \tau_k Q + (1 - \tau_k)y_k\} \\ &\geq \min_{y\in Q}\{\frac{L}{2}\|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1}\rangle\}.\end{aligned}$$

Finally, we conclude with

$$\begin{aligned}\Psi_{k+1}^* &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k \\ &\quad + B_{k+1}\min_{y\in Q}\{f_{k+1} + \langle g_{k+1}, y - x_{k+1}\rangle + \frac{L}{2}\|y - x_{k+1}\|_E^2\} \\ &\overset{(4.2),(6.3)}{\geq} (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}(f(w_{k+1}) - \delta) \\ &\overset{(6.5)}{\geq} A_{k+1}f(y_{k+1}) - E_k - B_{k+1}\delta = A_{k+1}f(y_{k+1}) - E_{k+1}.\end{aligned}$$

where $E_{k+1} = E_k + B_{k+1}\delta = \left(\sum_{i=0}^{k+1} B_i\right)\delta$. □

Using this lemma, we can obtain now the following convergence rate for the Intermediate Gradient Method:

Theorem 6.2. *Assume that the IGM is applied to a function f endowed with a (δ, L)-oracle with the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ satisfying (6.1) and (6.2). Then for all $k \geq 0$,we have*

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k} + \frac{\sum_{i=0}^k B_i}{A_k}\delta$$

where $A_k = \sum_{i=0}^k \alpha_i$.

Proof. Denote $f_i = f_{\delta,L}(x_i)$ and $g_i = g_{\delta,L}(x_i)$.
Then

$$\begin{aligned}\Psi_k^* &= \min_{x\in Q}\{Ld(x) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x - x_i\rangle]\} \\ &\leq Ld(x^*) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x^* - x_i\rangle]\end{aligned}$$

Using the Lemma 6.1, we have $A_k f(y_k) \leq Ld(x^*) + A_k f(x^*) + E_k$ i.e.

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k} + \frac{E_k}{A_k} = \frac{Ld(x^*)}{A_k} + \left(\frac{\sum_{i=0}^k B_i}{A_k}\right)\delta.$$

□

We conclude that

- ◇ The convergence rate of the IGM in the exact case is given by $\frac{Ld(x^*)}{A_k}$ and depends therefore only on $A_k = \sum_{i=0}^k \alpha_i$. When the oracle is exact, the sequence α_i must be chosen as large as possible i.e. growing linearly with i (corresponding to the condition $\alpha_i^2 = A_i$).

- ◇ The rate of error accumulation is given by $\frac{\sum_{i=0}^k B_i}{A_k}\delta$. At first sight, we have to choose $\{\alpha_i\}_{i\geq 0}$ as big as possible and $\{B_i\}_{i\geq 0}$ as small as possible. However the two sequences are linked by the constraint $\alpha_i^2 \leq B_i$. There is a trade-off to find between $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ depending on the level of the oracle error δ. We will come back to the choice of these two sequences in the following sections.

6.1.2 Variant with prox-type subproblems

The intermediate gradient method presented in the subsection 6.1.1 can be used with any norm $\|.\|_E$ and any prox-function d (which must be chosen such that $d(x^*)$ is small and subproblems based on d are easy). However, in this scheme, the subproblem $\min_{x\in Q}\{\langle g_{\delta,L}(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_E^2\}$ defining w_k is not based on the prox-function but on the squared norm. Such kind of subproblems can be difficult to solve and we consider in this section a variant of the intermediate gradient method which only use subproblems based on the prox-function and the corresponding Bregman distance. We propose the following modification of the IGM:

Algorithm 20 Intermediate Gradient Method (IGM)

1: Compute $x_0 = \arg\min_{x\in Q} d(x)$
2: Obtain $(f_{\delta,L}(x_0), g_{\delta,L}(x_0))$
3: Compute

$$y_0 = \arg\min_{x\in Q}\{Ld(x) + \alpha_0\langle g_{\delta,L}(x_0), x - x_0\rangle\} \tag{6.8}$$

4: **for** k = 0 : ... **do**
5: Compute

$$z_k = \arg\min_{x\in Q}\{Ld(x) + \sum_{i=0}^{k}\alpha_i\langle g_{\delta,L}(x_i), x - x_i\rangle\} \tag{6.9}$$

6: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k$
7: Obtain $(f_{\delta,L}(x_{k+1}), g_{\delta,L}(x_{k+1}))$
8: Compute

$$\hat{x}_{k+1} = \arg\min_{x\in Q}\{LV(x, z_k) + \alpha_{k+1}\langle g_{\delta,L}(x_{k+1}), x - z_k\rangle\} \tag{6.10}$$

9: Compute $w_{k+1} = \tau_k\hat{x}_{k+1} + (1 - \tau_k)y_k$.
10: Compute

$$y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}}y_k + \frac{B_{k+1}}{A_{k+1}}w_{k+1}. \tag{6.11}$$

11: **end for**

This method is more complicated but uses only subproblems based on the prox-function $d(x)$ (or equivalently on the corresponding Bregman distance $V(x, z)$). This property can be crucial in some situations. Furthermore, the convergence

rate is the same than for the basic IGM:

Lemma 6.3. *Assume that the modified IGM is applied to a function f endowed with a (δ, L)-oracle. For all $k \geq 0$, we have $A_k f(y_k) \leq \Psi_k^* + E_k$ where $E_k = \left(\sum_{i=0}^k B_i\right)\delta$ and $\Psi_k^* = \min_{x\in Q}\{\Psi_k(x) = Ld(x) + \sum_{i=0}^k \alpha_i[f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i\rangle]\}$.*

Proof. Denote $f_k = f_{\delta,L}(x_k)$ and $g_k = g_{\delta,L}(x_k)$.

- ◇ It is true for $k = 0$. Indeed

$$\begin{aligned}\Psi_0^* &\overset{(6.8)}{=} Ld(y_0) + \alpha_0[f_0\langle g_0, y_0 - x_0\rangle] \\ &\overset{(2.8)}{\geq} \alpha_0[f_0 + \langle g_0, y_0 - x_0\rangle + \frac{L}{2}\|y_0 - x_0\|_E^2] \\ &= \alpha_0 f(y_0) - \delta B_0.\end{aligned}$$

- ◇ If it is true for $k \geq 0$, it is also true for $k + 1$.
  Indeed, by the optimality condition of the subproblem defining z_k, we have

$$\langle L\nabla d(z_k) + \sum_{i=0}^k \alpha_i g_i, x - z_k\rangle \geq 0, \quad \forall x \in Q.$$

  Therefore

$$\begin{aligned}\Psi_{k+1}(x) &= Ld(x) + \sum_{i=0}^{k+1}\alpha_i[f_i + \langle g_i, x - x_i\rangle] \\ &\overset{(2.9)}{=} LV(x, z_k) + Ld(z_k) + \langle L\nabla d(z_k), x - z_k\rangle \\ &\quad + \sum_{i=0}^{k+1}\alpha_i[f_i + \langle g_i, x - x_i\rangle] \\ &\geq LV(x, z_k) + Ld(z_k) + \sum_{i=0}^{k}\alpha_i[f_i + \langle g_i, z_k - x_i\rangle] \\ &\quad + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ &\overset{(6.9)}{=} \Psi_k^* + LV(x, z_k) + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle].\end{aligned}$$

On the other hand, we have, assuming $\Psi_k^* \geq A_k f(y_k) - E_k$:

$$\begin{aligned}
& \Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
\geq\ & A_k f(y_k) - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
=\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + (B_{k+1} - \alpha_{k+1}) f(y_k) \\
& + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
\overset{(4.2)}{\geq}\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k \\
& + (B_{k+1} - \alpha_{k+1})[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] \\
& + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
=\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} \\
& + \langle g_{k+1}, (B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle.
\end{aligned}$$

As $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$ and $\tau_k = \frac{\alpha_{k+1}}{B_{k+1}}$, we have

$$(B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1}) = \alpha_{k+1}(x - z_k).$$

Therefore

$$\begin{aligned}
& \Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
\geq\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle.
\end{aligned}$$

We conclude that

$$\begin{aligned}
\Psi_{k+1}^* \ \geq\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} \\
& + \min_{x \in Q}\{L V(x, z_k) + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle\} \\
\overset{(6.10)}{=}\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} \\
& L V(\hat{x}_{k+1}, z_k) + \alpha_{k+1}\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle \\
=\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k \\
& + B_{k+1}[f_{k+1} + \frac{L}{B_{k+1}} V(\hat{x}_{k+1}, z_k) + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle] \\
\overset{(2.10)}{\geq}\ & (A_{k+1} - B_{k+1}) f(y_k) - E_k \\
& + B_{k+1}[f_{k+1} + \frac{L}{2B_{k+1}} \|\hat{x}_{k+1} - z_k\|_E^2 + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle].
\end{aligned}$$

As $\alpha_{k+1}^2 \leq B_{k+1}$, we have $\frac{1}{B_{k+1}} \geq \tau_k^2$ and therefore

$$\begin{aligned}
& \Psi_{k+1}^* \geq (A_{k+1} - B_{k+1}) f(y_k) - E_k \\
& + B_{k+1}[f_{k+1} + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle + \frac{\tau_k^2 L}{2} \|\hat{x}_{k+1} - z_k\|_E^2].
\end{aligned}$$

But $\tau_k(\hat{x}_{k+1} - z_k) = w_{k+1} - x_{k+1}$ and we obtain

$$\begin{aligned}\Psi^*_{k+1} &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k \\ &\quad + B_{k+1}[f_{k+1} + \langle g_{k+1}, w_{k+1} - x_{k+1}\rangle + \frac{L}{2}\|w_{k+1} - x_{k+1}\|^2_E] \\ &\overset{(4.2)}{\geq} (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}[f(w_{k+1}) - \delta] \\ &\overset{(6.11)}{\geq} A_{k+1}f(y_{k+1}) - E_k - B_{k+1}\delta.\end{aligned}$$

□

We can now apply the Theorem 6.2 to this modified IGM and conclude that it exhibits exactly the same convergence rate as the original IGM developed in subsection 6.1.1.

6.2 Link with existing methods

6.2.1 Link with Fast Gradient Method

If the sequence $\{B_k\}_{k\geq 0}$ is chosen such that $B_k = A_k$ for all $k \geq 0$, we have $y_k = w_k$ for all $k \geq 0$ and the IGM is nothing else than the scheme developed in [59] and described in subsection 2.4.5 of this thesis:

Algorithm 21 Fast Gradient Method (FGM)

1: Compute $x_0 = \arg\min_{x\in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$
4: Compute $y_k = \arg\min_{x\in Q}\{\frac{L}{2}\|x - x_k\|^2_E + \langle g_{\delta,L}(x_k), x - x_k\rangle\}$
5: Compute $z_k = \arg\min_{x\in Q}\{Ld(x) + \sum_{i=0}^{k}\alpha_i\langle g_{\delta,L}(x_i), x - x_i\rangle\}$
6: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k$.
7: **end for**

This method exhibits the following convergence rate (see Theorem 4.8 in Chapter 4)

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i}{A_k}\delta. \tag{6.12}$$

Whatever the choice made for α_i (and therefore A_i), the rate of error accumulation is of order $\Theta(k\delta)$. Therefore, the coefficients $\{\alpha_i\}_{i\geq 0}$ must be chosen as big as possible (since a smaller α_i would slow down the rate of convergence without reducing the rate of error accumulation) i.e. with a linear growth

$\alpha_i = \Theta(i)$. In [59], the choice $\alpha_i = \frac{i+1}{2}$ is suggested and it leads to a Fast Gradient Method (FGM) with an optimal convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$ in the exact case and rate of error accumulation $\Theta(k\delta)$ as we have established in Chapter 4.

Compared to this existing scheme developed in [59], the IGM offers an additional degree of freedom: we can choose B_k smaller than A_k. In this case, we replace $y_k = w_k$, by the more conservative rule $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$. With this modification:

1. we slow down the rate of errors accumulation $\frac{\sum_{i=0}^{k} B_i}{A_k} \leq \frac{\sum_{i=0}^{k} A_i}{A_k}$
2. we slow down the rate of convergence (this is unavoidable in view of Theorem 4.11) due to the condition $\alpha_k^2 \leq B_k$ (instead of $\alpha_k^2 \leq A_k$).

6.2.2 Link with Dual Gradient Method

If we choose constant coefficients $\alpha_i = 1$ and $B_i = 1$ in the IGM, we obtain the following scheme:

Algorithm 22 Dual Gradient Method (DGM)

Compute $x_0 = \arg\min_{x \in Q} d(x)$
for k = 0 : ... **do**
  Obtain $(f_{\delta,L}(x_k), g_{\delta,L}(x_k))$
  Compute $w_k = \arg\min_{x \in Q} \{ \frac{L}{2} \|x - x_k\|_E^2 + \langle g_{\delta,L}(x_k), x - x_k \rangle \}$
  Compute $y_k = \frac{1}{k} \sum_{i=0}^{k} w_i$
  Compute $x_{k+1} = \arg\min_{x \in Q} \{ Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle \}$
end for

This scheme is nothing else than the Dual Gradient Method (DGM) developed in [62] and described in subsection 2.4.4 of this thesis. This method exhibits the same behavior as the classical Gradient Method, with a slow convergence rate $\Theta(\frac{LR^2}{k})$ in the exact case and no accumulation of errors (see subsection 4.3.2): $f(y_k) - f^* \leq \frac{LR^2}{2(k+1)} + \delta$.

Within the family of Intermediate Gradient Methods, the Dual Gradient Method and the Fast Gradient Method can be seen as two extreme cases. The Dual Gradient method is slow (convergence rate $\Theta\left(\frac{LR^2}{k}\right)$ in the exact case) but robust with respect to oracle errors (no accumulation of errors, able to reach any accuracy bigger than δ). The Fast Gradient is fast (optimal convergence

rate $\Theta\left(\frac{LR^2}{k^2}\right)$ in the exact case) but highly sensitive with respect to oracle error (accumulation of errors at a linear rate $\Theta(k\delta)$ and unable to reach an accuracy better than $\epsilon^*_{FGM} = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$).

In the following sections, using our degrees of freedom for the choice of $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ in the IGM, we develop new methods with intermediate behavior between DGM and FGM:

- ◇ In Section 6.4, using a sequence of coefficients α_i that grows linearly (like in the FGM) before switching to a constant value (like in the DGM), we obtain a method that can be seen as a switching between FGM and DGM. The switching moment is optimized in order to reach a target accuracy ϵ in a minimal number of iterations.

- ◇ In Section 6.6, using a power policy $\alpha_i = \Theta(i^{p-1})$, we obtain methods with an intermediate convergence rate $\Theta\left(\frac{LR^2}{k^p}\right)$ $(1 \leq p \leq 2)$ and the corresponding intermediate (and optimal) rate of error accumulation $\Theta(k^{p-1}\delta)$.

6.3 Optimal choice of the coefficients for a target accuracy ϵ

6.3.1 Optimal Policy

We have developed a general family of first-order methods characterized by two sequences of coefficients, $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ that must satisfy the conditions $\alpha_i^2 \leq B_i \leq \sum_{j=0}^{i} \alpha_j$ and $0 \leq \alpha_i \leq B_i$ for all $i \geq 0$. For a given choice of $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$, we know that the accuracy obtained by the corresponding IGM method after k iterations is $f(y_k) - f^* \leq \frac{Ld(x^*)+\sum_{i=0}^{k} B_i\delta}{\sum_{i=0}^{k} \alpha_i}$. The behavior of an intermediate gradient method is directly governed by the choice of these coefficients.

In this section, we assume that we have an oracle with fixed accuracy δ and that we want to obtain a solution with target accuracy ϵ using a minimal number of iterations. Therefore, we are interested in an optimal policy for the coefficients $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$, i.e. in the choice of coefficients that minimizes the needed number of iteration to reach a final accuracy $\epsilon > \delta$. We denotes by $\theta := \frac{\epsilon}{\delta} > 1$, the ratio between target accuracy and oracle accuracy.

This optimal choice of the coefficients corresponds to the following optimization problem:

$$k^*(\delta,\theta) = \min_{\alpha\geq 0, B\geq 0, k\geq 0} k$$

such that

$$\frac{Ld(x^*)}{\delta} + \sum_{i=0}^{k} B_i \leq \theta \sum_{i=0}^{k} \alpha_i$$

$$\alpha_i^2 \leq B_i \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i = 0, ..., k$$

$$0 \leq \alpha_i \leq B_i, \quad \forall i = 0, ..., k.$$

Clearly in a optimal policy, we have $B_i = \max(\alpha_i, \alpha_i^2)$ and the optimal choice for $\{\alpha_i\}_{i\geq 0}$ is given by

$$k^*(\delta,\theta) = \min_{\alpha\in\Omega, k\geq 0} k$$

such that

$$\xi(k,\alpha) \geq 0$$

where $\xi(k,\alpha) = \theta \sum_{i=0}^{k} \alpha_i - \sum_{i=0}^{k} \max(\alpha_i, \alpha_i^2) - \frac{Ld(x^*)}{\delta}$ and $\Omega = \left\{\alpha \in \mathbb{R}_+^\infty | \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i \geq 0\right\}$.

Without loss of generality, we can assume that

- $\alpha_i \geq 1$ **for all** i.

  Indeed assume that we have a policy α such that $\alpha_l < 1$ for a given $l \geq 0$. Then, let us construct a new policy $\overline{\alpha}$ defined by $\overline{\alpha}_l = 1$ and $\overline{\alpha}_i = \alpha_i \ \forall i \neq l$. This new sequence $\overline{\alpha}$ is such that

  1. $\overline{\alpha} \in \Omega$
     Indeed:
     - for $i < l$: $\overline{\alpha_i}^2 = \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j = \sum_{j=0}^{i} \overline{\alpha}_j$
     - for $i = l$: $\overline{\alpha}_l^2 = 1 \leq 1 + \sum_{j=0}^{l-1} \alpha_j = \sum_{j=0}^{l} \overline{\alpha}_j$
     - for $i > l$: $\overline{\alpha_i}^2 = \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j \leq \sum_{j=0}^{i} \overline{\alpha}_j$.
  2. If $k \geq l$:

$$\begin{aligned}\xi(k,\overline{\alpha}) &= \theta \sum_{i=0}^{k} \alpha_i + \theta(1-\alpha_l) - \sum_{i=0}^{k} \max(\alpha_i, \alpha_i^2) \\ &\quad -(1-\alpha_l) - \frac{Ld(x^*)}{\delta} \\ &= \xi(k,\alpha) + (1-\alpha_l)(\theta - 1) > \xi(k,\alpha)\end{aligned}$$

and if $k < l$, $\xi(k, \overline{\alpha}) = \xi(k, \alpha)$.

Therefore

$$\min\{k|\xi(k, \overline{\alpha}) \geq 0\} \leq \min\{k|\xi(k, \alpha) \geq 0\}$$

and we conclude that the new policy $\overline{\alpha}$ is at least as good as α.

As we can assume w.l.o.g. that $\alpha_i \geq 1$ for all $i \geq 0$, we have $\max\{\alpha_i, \alpha_i^2\} = \alpha_i^2$ and our problem becomes

$$k^*(\delta, \theta) = \min_{\alpha \in \tilde{\Omega}, k} k$$

such that

$$\xi(k, \alpha) \geq 0$$

where $\xi(k, \alpha) = \theta \sum_{i=0}^{k} \alpha_i - \sum_{i=0}^{k} \alpha_i^2 - \frac{Ld(x^*)}{\delta}$ and $\tilde{\Omega} = \Omega \cap \{\alpha \in \mathbb{R}_+^\infty | \alpha_i \geq 1 \forall i \geq 0\}$.
In the following, we denote by $OptPol(\delta, \theta)$ this optimization problem defining the optimal choice for the sequence of coefficients $\{\alpha_i\}_{i \geq 0}$.

◇ **sequence α is increasing.**

Indeed, assume that we have a policy with $\alpha_{l+1} < \alpha_l$ for a given l satisfying $\min\{k \geq 0 | \xi(k, \alpha) \geq 0\} \geq l + 1$. Then, let us consider the new policy $\tilde{\alpha}$ defined by

$$\tilde{\alpha}_{l+1} = \alpha_l, \quad \tilde{\alpha}_l = \alpha_{l+1} \text{ and } \tilde{\alpha}_i = \alpha_i, \forall i < l \text{ or } i > l + 1.$$

This new policy is such that

1. $\tilde{\alpha} \in \tilde{\Omega}$
   Indeed, we have:
   - for $i < l$: $\tilde{\alpha}_i^2 = \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j = \sum_{j=0}^{i} \tilde{\alpha}_j$
   - for $i = l$:

$$\begin{aligned}
\tilde{\alpha}_l^2 &= \alpha_l^2 + (\tilde{\alpha}_l - \alpha_l)^2 + 2\alpha_l(\tilde{\alpha}_l - \alpha_l) \\
&\leq \sum_{j=0}^{l} \alpha_j + (\tilde{\alpha}_l - \alpha_l)^2 + 2\alpha_l(\tilde{\alpha}_l - \alpha_l) \\
&= \sum_{j=0}^{l} \tilde{\alpha}_j + (\alpha_l - \tilde{\alpha}_l)(1 - \alpha_l - \tilde{\alpha}_l) \\
&\leq \sum_{j=0}^{l} \tilde{\alpha}_j
\end{aligned}$$

- for $i = l+1$:

$$\tilde{\alpha}_{l+1}^2 = \alpha_l^2 \leq \sum_{j=0}^{l} \alpha_j \leq \sum_{k=0}^{l+1} \alpha_j = \sum_{k=0}^{l+1} \tilde{\alpha}_j$$

- for $i > l+1$: $\tilde{\alpha}_i^2 = \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_i = \sum_{j=0}^{i} \tilde{\alpha}_i$.

2. For all $k \geq l+1$: $\xi(k, \alpha) = \xi(k, \tilde{\alpha})$.

We conclude that the modified policy $\tilde{\alpha}$ is at least as good as the original one α (i.e. $\min\{k \geq 0 | \xi(k, \tilde{\alpha}) \geq 0\} \leq \min\{k \geq 0 | \xi(k, \alpha) \geq 0\}$) and we can therefore assume w.l.o.g. that an optimal policy is increasing.

In a feasible policy, the coefficients α_i cannot be as big as we want. The growth of this sequence is limited by the constraints $\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j$. In the exact case (i.e. when $\delta = 0$), it is clear that we have to choose the coefficients α_i as big as possible. In the inexact case, due to the accumulation of errors, this is not anymore always true. However, when the relative accuracy $\theta = \frac{\epsilon}{\delta}$ is sufficiently big, then it is possible to prove that the sequence defined by the recurrence

$$\hat{\alpha}_i = \frac{1 + \sqrt{1 + 4\sum_{j=0}^{i-1} \hat{\alpha}_j}}{2}$$

(i.e. $\hat{\alpha_i}^2 = \sum_{j=0}^{i} \hat{\alpha}_j$) with $\hat{\alpha}_0 = 1$, corresponding to the feasible policy with the highest coefficients, is still optimal even if $\delta \neq 0$.

Remark 6.3. For any policy feasible for the problem $OptPol(\delta, \theta)$, we have $\alpha_i \leq \hat{\alpha}_i, \quad \forall i = 0, ..., k$.

The following Theorem shows that under some particular conditions, the policy $\{\hat{\alpha}_i\}_{i \geq 0}$ is optimal for the general problem $OptPol(\delta, \theta)$.

Theorem 6.4. *If there exists $k \geq 0$ such that $\hat{\alpha}_k \leq \frac{\theta}{2}$ and $\xi(k, \hat{\alpha}) \geq 0$, then the policy $\{\hat{\alpha}_i\}_{i \geq 0}$ is optimal.*

Proof. Let $\{\alpha_i\}_{i \geq 0}$ be another feasible policy. Then we have, for all $i \leq k$

$$\begin{aligned} \xi(i, \hat{\alpha}) &= \xi(i, \alpha) + \sum_{j=0}^{i} [\theta(\hat{\alpha}_j - \hat{\alpha}_j) - (\hat{\alpha}_j^2 - \alpha_j^2)] \\ &= \xi(i, \alpha) + \sum_{j=0}^{i} (\hat{\alpha}_j - \alpha_j)(\theta - \alpha_j - \hat{\alpha}_j) \geq \xi(i, \alpha) \end{aligned}$$

since for all $j \leq i \leq k$, we have $\hat{\alpha}_j \geq \alpha_j$ and $\hat{\alpha}_j + \alpha_j \leq 2\hat{\alpha}_j \leq 2\hat{\alpha}_k \leq \theta$. We conclude that for any other feasible policy and any $i \leq k$, we have: $\xi(i, \alpha) \leq \xi(i, \hat{\alpha})$. Furthermore as $\min\{i \geq 0 | \xi(i, \hat{\alpha}) \geq 0\} \leq k$, we conclude that $\min\{i \geq 0 | \xi(i, \alpha) \geq 0\} \geq \min\{i \geq 0 | \xi(i, \hat{\alpha}) \geq 0\}$. The policy $\{\hat{\alpha}_i\}_{i \geq 0}$ is at least as good as any other feasible policy and is therefore optimal. □

6.3.2 Lower Complexity bound

If we drop the family of constraint

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i \geq 0 \tag{6.13}$$

in $OptPol(\delta, \theta)$, we obtain a much simpler optimization problem $OptPolRelax(\delta, \theta)$:

$$k_{relax}(\delta, \theta) = \min_{k \in \mathbb{N}, \alpha} k$$

such that:

$$\theta \sum_{i=0}^{k} \alpha_i \geq \sum_{i=0}^{k} \alpha_i^2 + \frac{Ld(x^*)}{\delta}$$

and

$$\alpha_i \geq 1, \quad \forall i \geq 0.$$

Since this problem is a relaxation of $OptPol(\delta, \theta)$, $k^*_{relax}(\delta, \theta)$ provides us a lower bound on $k^*(\delta, \theta)$, the number of iterations needed by an optimal IGM (i.e. using an optimal policy for α) for reaching the accuracy $\theta\delta$. For this relaxed problem, since it is homogeneous in the different coefficients $\{\alpha_i\}_{i \geq 0}$, we can assume without loss of generality that the sequence α is constant i.e. $\alpha_i = \alpha$ for all $i \geq 0$. Our problem $OptPolRelax(\delta, \theta)$ becomes:

$$k^*_{relax}(\delta, \theta) = \min_{k \in \mathbb{N}, \alpha} k$$

such that:

$$\theta(k+1)\alpha \geq (k+1)\alpha^2 + \frac{Ld(x^*)}{\delta} \text{ and } \alpha \geq 1$$

which is equivalent to:

$$\min_{\alpha \geq 1} \frac{Ld(x^*)}{\delta(\theta\alpha - \alpha^2)} - 1.$$

The optimal solution of this problem is given by $\alpha^* = \max(1, \frac{\theta}{2})$ and we conclude that:

1. If $\theta \leq 2$ then $\alpha^* = 1$ and $k^*_{relax}(\delta, \theta) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1 = \frac{Ld(x^*)}{\epsilon-\delta} - 1$. Furthermore as the choice $\alpha_i = 1$ for all $i \geq 0$ satisfies also the constraints (6.13), we conclude that this policy, corresponding to a dual gradient method, is optimal also for the non relaxed problem $OptPol(\delta, \theta)$. We have therefore $k^*(\delta, \theta) = k^*_{relax}(\delta, \theta) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1$ in the case $1 \leq \theta \leq 2$.

2. If $\theta > 2$ then $\alpha^* = \frac{\theta}{2}$ and $k^*_{relax}(\delta, \theta) = \frac{4Ld(x^*)}{\delta\theta^2} - 1$. However, in this case the constant policy $\alpha_i = \frac{\theta}{2}$ does not satisfy the constraints (6.13) and we can only conclude that $k^*(\delta, \theta) \geq k^*_{relax}(\delta, \theta) = \frac{4Ld(x^*)}{\delta\theta^2} - 1$.

6.4 Switching policy for the coefficients

In the previous section, we have considered the problem of optimal choice of a policy from all the policies feasible for the Intermediate Gradient Method. With $\xi(k,\alpha) = \theta \sum_{i=0}^{k} \alpha_i - \sum_{i=0}^{k} \alpha_i^2 - \frac{Ld(x^*)}{\delta}$, this general problem can be expressed as:

$$k^*(\delta,\theta) = \min_{\alpha} k \tag{6.14}$$

such that

$$\xi(k,\alpha) \geq 0, \quad \alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j \text{ and } \alpha_i \geq 1, \quad \forall i = 0,...,k.$$

In this section, in order to be able to compute an optimal policy analytically, we restrict ourself to switching policies, a subclass of feasible policies for the IGM. More precisely, we consider policies of the form:

$$\alpha_i = \begin{cases} \frac{i+2}{2} & \text{when } i = 0,...,m\ , \\ l & \text{when } i = m+1,...,k \end{cases}$$

where:

- $k \geq 0$ denotes the total number of iterations that we perform
- $m \in \{0, k\}$ denotes the number of fast-gradient type iterations that we perform before to switch
- $l \geq 0$ denotes the level of the coefficients after the switching.

The motivations for this choice of coefficients are the following:

1. This policy is simple and can be easily interpreted. Using a switching policy, the IGM can be seen as a smart switching between FGM and DGM. In the case $m = 0$, $l = 1$, we retrieve the Dual Gradient Method and in the case $m = k$, we obtain a Fast Gradient Method. In between, when $0 < m < k$, we start with coefficients growing linearly like in the Fast Gradient Method before switching to constant coefficients like in the Dual Gradient Method. This is not a pure switching, since after m iterations of the FGM, we do not start the DGM from scratch using as initial iterate the last iterate obtained by the FGM. Instead, the first-order information obtained before the switching stays in the model $\Psi_i(x)$ $(m \leq i \leq k)$ with their linearly growing coefficients but the new first-order information, obtained after the switching, enters the model with constant coefficients like in the DGM.

2. In spite of its simplicity, the optimal switching policy leads, as we will see in subsection 6.4.4, to a complexity of the same order than what could be expected using the general optimal policy. We lose nothing (except possibly some small constant factor in the complexity) with the restriction to switching policies.

6.4.1 Feasible Policy

This switching policy is feasible for the IGM iff

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_i, \quad \forall i = 0, ..., k \tag{6.15}$$

and

$$\alpha_i \geq 1, \quad \forall i = 0, ..., k. \tag{6.16}$$

When $i \leq m$, the condition (6.15) gives us $\frac{(i+2)^2}{4} \leq \frac{1}{4}(i+1)(i+4)$ which is satisfied for every $i \geq 0$.
For $i = m+1$, the condition (6.15) is $l^2 \leq \sum_{j=0}^{m} \frac{j+2}{2} + l$ which is satisfied by a positive l iff $l \leq \frac{\sqrt{m^2+5m+5}+1}{2}$.
For $i > m+1$, the condition (6.15) is trivially satisfied provided that it is satisfied for $i = m+1$.
On the other hand, the condition (6.16) is completely equivalent to $l \geq 1$. We conclude that our switching policy is feasible if and only if l satisfies

$$1 \leq l \leq \frac{\sqrt{m^2+5m+5}+1}{2}.$$

As $\frac{\sqrt{m^2+5m+5}+1}{2} = \Theta(m)$, for the simplicity of our analysis, we will use the easier (but stronger) condition:

$$1 \leq l \leq \frac{m+2}{2}.$$

Furthermore, we have also to take into account the implicit constraint $m \leq k$. We conclude that the optimization problem given the optimal switching policy becomes:

$$k^*_{Switch}(\delta, \theta) = \min_{k \in \mathbb{N}, m \in \mathbb{N}, l} k \tag{6.17}$$

such that

$$k \geq m \tag{6.18}$$

$$\xi(k, \alpha) \geq 0 \tag{6.19}$$

$$\alpha_i = \frac{i+2}{2}, \quad \forall i = 0, ..., m \tag{6.20}$$

$$\alpha_i = l, \quad \forall i = m+1, ..., k \tag{6.21}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{6.22}$$

We denote this optimization problem by $OptSwitchPol(\delta, \theta)$. As any feasible policy for $OptSwitchPol(\delta, \theta)$ is also feasible for $OptPol(\delta, \theta)$, we have that $k^*_{Switch}(\delta, \theta) \geq k^*(\delta, \theta)$.
As

$$\sum_{i=0}^{k} \alpha_i = \sum_{i=0}^{m} \frac{i+2}{2} + \sum_{i=m+1}^{k} l = \frac{1}{4}(m+1)(m+4) + (k-m)l$$

and

$$\begin{aligned} \sum_{i=0}^{k} \alpha_i^2 &= \sum_{i=0}^{m} \left(\frac{i+2}{2}\right)^2 + \sum_{i=m+1}^{k} l^2 \\ &= \frac{1}{24}(m+1)(2m^2+13m+24) + (k-m)l^2 \end{aligned}$$

we have

$$\xi(k, \alpha) = (k-m)(\theta l - l^2) - \frac{1}{24}(m+1)(2m^2+13m+24-6m\theta-24\theta) - \frac{Ld(x^*)}{\delta}$$

and the problem $OptSwitchPol(\delta, \theta)$ becomes:

$$k^*_{Switch}(\delta, \theta) = \min_{(k,m,l) \in \mathbb{R}^3_+} k \tag{6.23}$$

such that

$$k \geq m. \tag{6.24}$$

$$(k-m)(\theta l - l^2) \geq \frac{1}{24}(m+1)(2m^2+13m+24-6m\theta-24\theta) + \frac{Ld(x^*)}{\delta} \tag{6.25}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{6.26}$$

Let us consider two cases:

1. $\theta l - l^2 \leq 0$ **i.e.** $l \geq \theta$
   In this case, if (k, m, l) is feasible i.e. satisfies (6.24), (6.25) and (6.26) then (m, m, l) also satisfies these constraints with a better value of the objective function. We conclude that if in a optimal solution $l \geq \theta$ then necessarily $k = m$ and we have a FGM. But for a FGM, the value of l does not play any role and we can therefore assume without loss of generality that $l < \theta$.

2. $\theta l - l^2 > 0$ **i.e.** $l < \theta$
   In this case, our problem $OptSwitchPol(\delta, \theta)$ becomes:

$$k^*_{Switch}(\delta, \theta) = \min_{k \in \mathbb{N}, m \in \mathbb{N}, l} k$$

such that

$$k \geq m \tag{6.27}$$

$$k \geq m + \frac{Ld(x^*)}{\delta(\theta l - l^2)} + \frac{(m+1)(2m^2 + 13m + 24 - 6\theta m - 24\theta)}{24(\theta l - l^2)} \tag{6.28}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{6.29}$$

For a given choice of l and m (and dropping the integrality assumption on k), the needed number of iteration is given by

$$k = m + \frac{1}{\theta l - l^2} \max(0, \frac{Ld(x^*)}{\delta} + \frac{1}{24}(m+1)(2m^2 + 13m + 24 - 6\theta m - 24\theta)). \tag{6.30}$$

6.4.2 Optimal Switching Level l

For a given value of m, let us look to the optimal choice for the switching level l. In view of (6.30), this optimal switching $l^*(m)$ corresponds to the optimal solution of the following optimization problem in l:

$$\max_{l \in \mathbb{R}^+} \theta l - l^2 \text{ such that } 1 \leq l \leq \frac{m+2}{2}.$$

The function $\theta l - l^2$ is increasing before reaching its maximum at $l = \frac{\theta}{2}$ and decreasing after it. We conclude that the optimal choice for the switching level is $l^*(m) = 1$ if $\theta \leq 2$ and $l^*(m) = \min\left(\frac{m+2}{2}, \frac{\theta}{2}\right)$ if $\theta \geq 2$.

6.4.3 Optimal Switching Moment m

Using the optimal switching level i.e. $l = 1$ if $\theta \leq 2$ and $l = \min\left(\frac{m+2}{2}, \frac{\theta}{2}\right)$ if $\theta \geq 2$, the optimal choice for the switching moment m is given by the optimization problem

$$\min[M(m) := \max(m, F(m))]$$

where if $\theta \leq 2$ (and therefore $l = 1$)

$$F(m) = m + \frac{1}{\theta - 1}\left(\frac{Ld(x^*)}{\delta} + \frac{(m+1)}{24}(2m^2 + 13m + 24 - 6\theta m - 24\theta)\right)$$

and if $\theta \geq 2$:

$$\begin{aligned} F(m) &= m + \frac{4}{\theta^2}\left(\frac{Ld(x^*)}{\delta} + \frac{(m+1)}{24}(2m^2 + 13m + 24 - 6m\theta - 24\theta)\right) \\ &\quad \text{when } m \geq \theta - 2 \text{ (i.e. } l = \frac{\theta}{2}) \\ &= m + \frac{4}{(m+2)(2\theta - m - 2)}\left(\frac{Ld(x^*)}{\delta}\right. \\ &\quad \left. + \frac{(m+1)}{24}(2m^2 + 13m + 24 - 6m\theta - 24\theta)\right) \\ &\quad \text{when } m \leq \theta - 2 \text{ (i.e. } l = \frac{m+2}{2}). \end{aligned}$$

Let us start with the case $\theta \leq 2$.

Case 1 : $\theta \leq 2$

Theorem 6.5. *If $\theta \leq 2$, the optimal switching policy is given by $\alpha_i = 1$ for all $i \geq 0$ for which the Intermediate Gradient Method is nothing else that the Dual Gradient Method.*
The corresponding required number of iterations is given by

$$k^*_{Switch}(\delta, \theta) = \frac{Ld(x^*)}{\delta(\theta - 1)} - 1.$$

Proof. When $\theta \leq 2$, the function $F(m)$ is increasing in m on $\mathbb{R}_+$ and the minimizer of $M(m) = \max(m, F(m))$ on $\mathbb{N}$ is therefore $m = 0$.
This optimal number of fast iteration $m = 0$ corresponds to the policy $\alpha_i = 1$ for all $i \geq 0$.
The needed number of iterations is given by $M(0) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1$. □

Case 2: $\theta \geq 2$

When $\theta \geq 2$, the function $F(m)$ (and therefore also the function $M(m)$) is defined differently on the two intervals $[0, \theta-2]$ and $[\theta-2, +\infty[$. On $[\theta-2, +\infty[$, the minimum of M is easy to find:

Lemma 6.6. *Assume that $\theta \geq 2$ then we have* $\arg\min_{m \geq \theta-2} M(m) = \theta - 2$.

Proof. When $\theta \geq 2$ and $m \geq \theta - 2$, we have $F'(m) = \frac{6m^2+(30-12\theta)m+37-30\theta+6\theta^2}{6\theta^2} > 0$. As m and $F(m)$ are increasing function on $[\theta-2, +\infty[$, we have that $M(m)$ is also increasing on this interval and therefore $\arg\min_{m \geq \theta-2} M(m) = \theta - 2$. □

Remark 6.4. We assume for the rest of the chapter that the desired final accuracy can be bounded by $\epsilon \leq Ld(x^*) + \delta$. This assumption is natural. Indeed instead if we had $\frac{Ld(x^*)}{\epsilon-\delta} < 1$, it would mean that our problem with oracle accuracy δ could be solved up to target accuracy ϵ in one iteration by the Gradient Method, and is therefore completely trivial.

On the interval $[0, \theta-2]$, the situation is more complicated. Two cases are possible, depending on the relative position of θ compared to a threshold θ_r which is defined as the unique root of

$$R(\theta) := \frac{2\theta^3}{3} + \frac{\theta^2}{2} - \frac{13\theta}{6} + 1 - \frac{4Ld(x^*)}{\delta}$$

greater than 2.

Remark 6.5. This polynomial has one and only one root greater than 2. Indeed, we have:

- $R(2) = 4\left(1 - \frac{Ld(x^*)}{\delta}\right) \leq 0$ (since $Ld(x^*) \geq \epsilon - \delta \geq \delta$)
- $\lim_{\theta \to +\infty} R(\theta) = +\infty$
- $R'(\theta) > 0$ for all $\theta \geq 1$.

First, let us prove the following lemma that gives another characterization of the conditions $\theta \geq \theta_r$:

Lemma 6.7. *Let us define the function:*

$$N(m) = \frac{4Ld(x^*)}{\delta} + \frac{(m+1)}{6}(24 + 13m + 2m^2 - 6m\theta - 24\theta).$$

The condition $\theta \geq \theta_r$ is completely equivalent with the existence of a root for $N(.)$ on $[0, \theta-2]$.

Proof. The function $N(.)$ is such that:

- $N(0) \geq 0$ since $\theta \leq \frac{Ld(x^*)}{\delta} + 1$ by assumption
- $N'(m) = m^2 + (5 - 2\theta)m + \frac{37}{6} - 5\theta$ is strictly negative on$]\theta - \frac{5}{2} - \sqrt{\frac{1}{12} + \theta^2}, \theta - \frac{5}{2} + \sqrt{\frac{1}{12} + \theta^2}[$ an therefore also on $[0, \theta - 2]$.

Therefore, N has a root on $[0, \theta - 2]$ iff $N(\theta - 2) \leq 0$.

But we have that $N(\theta - 2) \leq 0$ is equivalent with $R(\theta) \geq 0$. The function $R(.)$ is such that $R(2) = 4\left(1 - \frac{Ld(x^*)}{\delta}\right) \leq 0$ (since $Ld(x^*) \geq \epsilon - \delta \geq \delta$) and $R'(\theta) > 0$ for all $\theta \geq 1$. Therefore:

$$R(\theta) \geq 0 \Leftrightarrow \theta \geq \theta_r.$$

We conclude that the existence of a root for $N(m)$ between 0 and $\theta - 2$ is completely equivalent with the condition $\theta \geq \theta_r$. □

In the same way, we can also prove the following equivalence:

Lemma 6.8. *The condition $\theta \leq \theta_r$ is completely equivalent with $N(m) \geq 0$ for all $m \in [0, \theta - 2]$.*

Proof. As $N(0) \geq 0$ and $N'(m) < 0$ on $[0, \theta - 2]$, we have that $N(m) \geq 0$ on this interval iff $N(\theta - 2) \geq 0$.
But we have that the condition $N(\theta - 2) \geq 0$ is equivalent with $R(\theta) \leq 0$. As $R(2) \leq 0$ and $R'(\theta) > 0$ for all $\theta \geq 1$, this last condition is itself equivalent with $\theta \leq \theta_r$. □

The following lemma provides us with an estimation of the threshold θ_r:

Lemma 6.9. *The threshold θ_r is such that:*

$$\theta_r \in \left[2\sqrt[3]{\frac{5}{7}\frac{Ld(x^*)}{\delta}}, 2\sqrt[3]{\frac{Ld(x^*)}{\delta}}\right] = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right).$$

Proof.
- For all $\theta \geq 2$, we have $\frac{\theta^2}{2} - \frac{13\theta}{6} + 1 \geq \alpha\theta^3$ with $\alpha = \min_{\theta \geq 2}\left(\frac{\frac{\theta^2}{2} - \frac{13\theta}{6} + 1}{\theta^3}\right) = \frac{-1}{6}$. We conclude that $\frac{4Ld(x^*)}{\delta} = \frac{2\theta_r^3}{3} + \frac{\theta_r^2}{2} - \frac{13\theta_r}{6} + 1 \geq \frac{1}{2}\theta_r^3$ and therefore $\theta_r \leq 2\sqrt[3]{\frac{Ld(x^*)}{\delta}}$.
- For all $\theta \geq 2$, we have $\frac{\theta^2}{2} - \frac{13\theta}{6} + 1 \leq \beta\theta^3$ with $\beta = \max_{\theta \geq 2}\left(\frac{\frac{\theta^2}{2} - \frac{13\theta}{6} + 1}{\theta^3}\right) \leq 0.03061..$ We conclude that $\frac{4Ld(x^*)}{\delta} = \frac{2\theta_r^3}{3} + \frac{\theta_r^2}{2} - \frac{13\theta_r}{6} + 1 \leq \frac{7}{10}\theta_r^3$ and therefore $\theta_r \geq 2\sqrt[3]{\frac{5}{7}\frac{Ld(x^*)}{\delta}}$.

□

Now we are able to study the behavior of $M(.)$ on the interval $[0, \theta - 2]$. We have to consider two subcases:

Case 2.1: $2 \leq \theta \leq \theta_r$
For the simplicity of the analysis, we assume here that the relative desired accuracy $\theta = \frac{\epsilon}{\delta}$ is an integer.

Lemma 6.10. *Assume that θ is an integer on $\{2, \lfloor \theta_r \rfloor\}$ then*

$$\arg \min_{m \in \{0, \theta-2\}} M(m) = \theta - 2.$$

Proof. We have: $F(m) = m + \frac{N(m)}{D(m)}$ with $D(m) = (m+2)(2\theta - m - 2)$. First, let us establish some useful properties of the functions $N(m)$ and $D(m)$.

1. $N(m) \geq 0$ for all $m \leq \theta - 2$ using the fact that $\theta \leq \theta_r$ and the lemma 6.8.
2. $D(m) > 0$ on $]-2, 2\theta - 2[$ and therefore also on $[0, \theta - 2]$
3. $N'(m) = m^2 + (5 - 2\theta)m + \frac{37}{6} - 5\theta$ is strictly negative on$]\theta - \frac{5}{2} - \sqrt{\frac{1}{12} + \theta^2}, \theta - \frac{5}{2} + \sqrt{\frac{1}{12} + \theta^2}[$ an therefore also on $[0, \theta - 2]$.
4. $D'(m) = -2m + 2\theta - 4$ is strictly positive on $[0, \theta - 2[$ and $D'(\theta - 2) = 0$.

Since $N(m)$ is positive and $D(m)$ is strictly positive on $[0, \theta - 2]$, we have that $F(m) \geq m$ on this interval and therefore $M(m) = F(m)$ for all $m \leq \theta - 2$. Now, let us prove that $F'(m) < 0$ for all $m \in [0, \theta - \frac{13}{6}]$. Indeed, let $m \in [0, \theta - \frac{13}{6}]$ we have:

$$\begin{aligned} & F'(m) < 0 \\ \Leftrightarrow \quad & D^2(m) + N'(m)D(m) < D'(m)N(m) \\ \Leftrightarrow \quad & D(m) + N'(m) < D'(m)\frac{N(m)}{D(m)} \end{aligned}$$

where the last equivalence comes from the fact that $D(m) > 0$. Now the last inequality is satisfied since $D'(m) > 0$, $N(m) \geq 0$, $D(m) > 0$ and $D(m) + N'(m) = m - \theta + \frac{13}{6} \leq 0$.

We conclude that $F'(m) < 0$ for all $m \in [0, \theta - \frac{13}{6}]$. Therefore, since $\theta \in \mathbb{N}$, the minimizer of $M(.)$ on $\{0, \theta - 2\}$ can be $\theta - 2$ or $\theta - 3$.

But we have that

$$M(\theta-3) - M(\theta-2) = F(\theta-3) - F(\theta-2) = \frac{-\frac{2\theta^3}{3} + \frac{\theta^2}{2} + \frac{13\theta}{6} + \frac{4Ld(x^*)}{\delta} - 1}{\theta^2(\theta^2 - 1)} \geq 0$$

since $R(\theta) \geq 0$ (i.e. $\theta \leq \theta_r$).

We conclude that $\arg\min_{m\in\{0,\theta-2\}} M(m) = \arg\min_{m\in\{0,\theta-2\}} F(m) = \theta - 2$. □

We are now able to obtain the optimal switching policy in the case $2 \leq \theta \leq \theta_r$ (adding however an integer assumption on θ):

Theorem 6.11. *Assume that $\theta \in \{2, \lfloor \theta_r \rfloor\}$ then the optimal switching policy is given by*

$$\alpha_i = \begin{cases} \frac{i+2}{2} & \text{when } i \leq \theta - 2\,, \\ \frac{\theta}{2} & \text{when } i \geq \theta - 2\,. \end{cases}$$

The corresponding needed number of iteration is given by

$$k^*_{Switch}(\delta, \theta) = \frac{4Ld(x^*)}{\delta\theta^2} + \frac{1}{\theta^2}\left(\frac{\theta^3}{3} - \frac{5\theta^2}{2} + \frac{13\theta}{6} - 1\right)$$

which belongs to

$$\left[\frac{4Ld(x^*)}{\delta\theta^2} - 1, \frac{6Ld(x^*)}{\delta\theta^2}\right].$$

Proof. By lemmas 6.6 and 6.10, we know that $\arg\min_{m\in\mathbb{N}} M(m) = \theta - 2$ i.e. that the optimal number of fast iterations is given by $\theta - 2$. By subsection 6.4.2, we know that the optimal switching level is $l = \min\left(\frac{m+2}{2}, \frac{\theta}{2}\right) = \frac{\theta}{2}$. Therefore, the optimal switching policy is:

$$\alpha_i = \begin{cases} \frac{i+2}{2} & \text{when } i \leq \theta - 2\,, \\ \frac{\theta}{2} & \text{when } i \geq \theta - 2\,. \end{cases}$$

The corresponding needed number of iterations is given by:

$$k^*_{Switch}(\delta, \theta) = M(\theta - 2) = \frac{4Ld(x^*)}{\delta\theta^2} + \frac{1}{\theta^2}\left(\frac{\theta^3}{3} - \frac{5\theta^2}{2} + \frac{13\theta}{6} - 1\right)$$

iterations. But as $\theta \leq \theta_r$, we have $R(\theta) \leq 0$ i.e. $\frac{\theta^3}{3} + \frac{\theta^2}{4} - \frac{13\theta}{12} + \frac{1}{2} \leq \frac{2Ld(x^*)}{\delta}$ which implies $\frac{\theta^3}{3} - \frac{5\theta^2}{2} + \frac{13\theta}{6} - 1 \leq \frac{2Ld(x^*)}{\delta}$ and therefore $k^*_{Switch}(\delta, \theta) \leq \frac{6Ld(x^*)}{\delta\theta^2}$. Furthermore as $\theta \geq 2$, $\frac{1}{\theta^2}\left(\frac{\theta^3}{3} - 5\frac{\theta^2}{2} + \frac{13\theta}{6} - 1\right) \geq -1$ and therefore $k^*_{Switch} \geq \frac{4Ld(x^*)}{\delta\theta^2} - 1$. We conclude that $k^*_{Switch}(\delta, \theta) \in [\frac{4Ld(x^*)}{\delta\theta^2} - 1, \frac{6Ld(x^*)}{\delta\theta^2}]$. □

Remark 6.6. During the intermediate regime (i.e. $2 \leq \theta \leq \theta_r$), we have:

$$\frac{\theta^2(\theta-2)\delta}{6Ld(x^*)} \leq \frac{m}{k} \leq \frac{\theta^2(\theta-2)\delta}{4Ld(x^*) - \delta\theta^2}.$$

The proportion of fast iterations $\frac{m}{k}$ grows from 0 to 1 with a rate proportional to θ^3.

Case 2.2: $\theta \geq \theta_r$

Lemma 6.12. *Assume that $\theta \geq \theta_r$ then $\arg\min_{m\geq 0} M(m) = \overline{m}$ where $\overline{m}$ is the unique root of $N(.)$ on $[0, \theta - 2]$.*

Proof. Since $\theta \geq \theta_r$, in view of lemma 6.7, there exist $\overline{m} \in [0, \theta - 2]$ such that $N(\overline{m}) = 0$ i.e. $F(\overline{m}) = \overline{m}$. As $N'(m) < 0$ for all $m \in [0, \theta - 2]$, we have that $N(m) > 0$ for all $m < \overline{m}$ and using the same reasoning that in the proof of lemma 6.10, we conclude that $F'(m) < 0$ for all $m < \overline{m}$. Therefore $\arg\min_{m\geq 0} M(m) = \overline{m}$ since

- $\diamond$ For all $m \geq \overline{m}$: $M(m) = Max(m, F(m)) \geq \overline{m} = M(\overline{m})$
- $\diamond$ For all $m \leq \overline{m}$: $M(m) = Max(m, F(m)) = F(m) \geq F(\overline{m}) = M(\overline{m})$.

□

We can therefore obtain the optimal switching policy in the case $\theta \geq \theta_r$:

Theorem 6.13. *Assume that $\theta \geq \theta_r$ then the optimal switching policy is $\alpha_i = \frac{i+2}{2}$ for all $i \geq 0$, for which the Intermediate Gradient Method is nothing else than a Fast Gradient Method.*
The corresponding needed number of iterations $k^*_{Switch}(\delta, \theta)$ is given by the unique root of $N(m) = \frac{4Ld(x^*)}{\delta} + \frac{(m+1)}{6}(24 + 13m + 2m^2 - 6m\theta - 24\theta)$ on $[0, \theta - 2]$. Furthermore, we have that

$$k^*_{Switch}(\delta, \theta) \in \left[2\sqrt{\frac{Ld(x^*)}{(\theta-1)\delta}} - 4, 2\sqrt{\frac{2Ld(x^*)}{(\theta-2)\delta}}\right] = \Theta\left(\sqrt{\frac{LR^2}{\theta\delta}}\right).$$

Proof. In view of lemma 6.12, the optimal switching moment is given by $\overline{m}$, the unique root of $N(.)$ on this interval, for which we have $M(\overline{m}) = F(\overline{m}) = \overline{m}$.

Since $M(\overline{m}) = \overline{m}$, it means that we have only to perform fast iterations: $\alpha_i = \frac{i+2}{2}$ for all $i \geq 0$ and that the needed number of iterations corresponds exactly to this root $\overline{m} = k^*_{Switch}(\delta, \theta)$.

Let us try now to obtain lower and upper bound for this quantity.

With $\alpha_i = \frac{i+2}{2}$ for all i, the convergence rate of the IGM (which is nothing else that a FGM) is given by:

$$f(y_k) - f^* \leq Acc_{FGM}(k) = \frac{Ld(x^*) + \frac{1}{24}(k+1)(2k^2 + 13k + 24)\delta}{\frac{1}{4}(k+1)(k+4)}.$$

The needed number of iteration $k^*_{Switch}(\delta,\theta) = k_{FGM}(\delta,\theta)$ corresponds to the first positive k such that $Acc_{FGM}(k) = \theta\delta$ (we drop here the integer assumption for k).
We have that $Acc_{FGM}(k) = \frac{4Ld(x^*)}{(k+1)(k+4)} + \frac{k+4}{4}\delta + \frac{1}{6}\frac{(k+1)k}{k+4}\delta$. Therefore:

1. **Upper-bound**
   $\overline{Acc}(k) := \frac{4Ld(x^*)}{(k+1)(k+4)} + \frac{(k+4)\delta}{2} \geq Acc(k)$ for any $k \geq 0$ and therefore $\overline{Acc}(k_{FGM}(\delta,\theta)) \geq \theta\delta$. This last inequality is equivalent with

   $$\frac{4Ld(x^*)}{(k_{FGM}(\delta,\theta)+1)(k_{FGM}(\delta,\theta)+4)} \geq (\theta - \frac{k_{FGM}(\delta,\theta)+4}{2})\delta$$

   and as $k_{FGM}(\delta,\theta) \leq \theta - 2$ (i.e. $\theta - \frac{k_{FGM}(\delta,\theta)+4}{2} \geq \frac{\theta}{2} - 1$), it implies that $\frac{4LR^2}{k_{FGM}(\delta,\theta)^2} \geq (\frac{\theta}{2} - 1)$ i.e.

   $$k_{FGM}(\delta,\theta) \leq 2\sqrt{\frac{2Ld(x^*)}{(\theta-2)\delta}}.$$

2. **Lower-bound**
   $\underline{Acc}(k) := \frac{4Ld(x^*)}{(k+1)(k+4)} + \frac{(k+4)}{4}\delta \leq Acc(k)$ and therefore $\underline{Acc}(k_{FGM}(\delta,\theta)) \leq \theta\delta$. This last inequality is equivalent with: $\frac{4Ld(x^*)}{(k_{FGM}(\delta,\theta)+1)(k_{FGM}(\delta,\theta)+4)} \leq (\theta - \frac{k_{FGM}(\delta,\theta)+4}{4})\delta$ which implies $\frac{4Ld(x^*)}{(k_{FGM}(\delta,\theta)+4)^2} \leq (\theta-1)\delta$ i.e.

   $$2\sqrt{\frac{Ld(x^*)}{(\theta-1)\delta}} - 4 \leq k_{FGM}(\delta,\theta).$$

We conclude that in the case $\theta \geq \theta_r$, $k_{Switch}(\delta,\theta) = k_{FGM}(\delta,\theta) = \Theta\left(\sqrt{\frac{LR^2}{\theta\delta}}\right)$. □

Remark 6.7. The FGM that we obtain here in the case $\theta \geq \theta_r$ is not completely equivalent with the original method developed in [59] whose behavior we have studied in the inexact case in Chapter 4. The original FGM corresponds to the IGM with $\alpha_i = \frac{i+1}{2}$ and $B_i = A_i = \sum_{j=0}^{i} \alpha_j$ for all $i \geq 0$.

With this classical choice, the method can be expressed using only three sequences $\{w_i\}_{i\geq 0}$, $\{z_i\}_{i\geq 0}$ and $\{x_i\}_{i\geq 0}$ since $A_i = B_i$ and therefore $y_i = w_i$ for all $i \geq 0$. The convergence rate is given by:

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)(k+2)} + \frac{2k+6}{6}\delta.$$

However, using the choice $\alpha_i = \frac{i+2}{2}$ for all $i \geq 0$ and the new degree of freedom given by the IGM to choose $B_i \neq A_i$, it is possible to improve slightly the FGM.

Indeed, using the additional sequence $y_i = \frac{A_i - B_i}{A_i} + \frac{B_i}{A_i}$ with $B_i = \alpha_i^2$, we obtain the convergence rate:

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)(k+4)} + \frac{2k^2 + 13k + 24}{6(k+4)}\delta$$

We conclude that the FGM considered here exhibits, at the price of an additional sequence, a slightly better convergence rate in the exact case and a slightly smaller accumulation of error.

Remark 6.8. When $2 \leq \theta \leq \theta_r$, we have $k^*_{Switch}(\delta, \theta) = \Theta\left(\frac{LR^2}{\delta\theta^2}\right)$ and when $\theta \geq \theta_r$, we have $k^*_{Switch}(\delta, \theta) = \Theta\left(\sqrt{\frac{LR^2}{\theta\delta}}\right)$. As $\theta_r = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$, we have $k^*_{Switch}(\delta, \theta_r) = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ and the transition in $\theta = \theta_r$ is continuous.

Remark 6.9. The second threshold θ_r does not correspond exactly to the best relative accuracy reachable by the fast gradient method $\theta^*_{FGM} = \frac{\epsilon^*_{FGM}}{\delta}$ (θ_r and θ^*_{FGM} both depending on L, R and δ). We have $\theta^*_{FGM} \leq \theta_r$ but for $\theta^*_{FGM} \leq \theta < \theta_r$, even if the accuracy θ can be reached by the FGM, it is better to use an intermediate method with switching after $m = \theta - 2 < k$.

However, we have that θ^*_{FGM} and θ_r are of the same order $\Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$. Therefore the numbers of iterations needed by the FGM and the best IGM for reaching relative accuracy θ are of the same order $\Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$. In particular, we see that the proportion of fast step in the optimal IGM is close to 1 (since $m = \theta - 2 = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$).

6.4.4 General optimality of the optimal switching policy

Now that we have obtained the optimal switching policy, let us compare the complexity of this policy $k^*_{Swicth}(\delta, \theta)$ with the complexity of a general optimal policy $k^*(\delta, \theta)$ (i.e. without the restriction to switching policies). The goal is of course to see if we lost something with the switching policies or if we could make this restriction without loss of generality.

We have to consider three different cases:

1. **When $\theta \leq 2$**:
   The optimal policy and the optimal switching policy coincide. The optimal IGM is nothing else than the Dual Gradient method i.e. with all coefficients α_i equals to one and a corresponding needed number of iterations given by $k^*(\delta,\theta) = k^*_{Switch}(\delta,\theta) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1$.

2. **When $2 \leq \theta \leq \theta_{\mathbf{r}}$**:
   As $2 \leq \theta$, in view of subsection 6.3.2, we know that an optimal policy for the general optimization problem $OptPol(\delta,\theta)$ cannot have a better complexity than $\frac{4Ld(x^*)}{\delta\theta^2} - 1$ (i.e. that $k^*(\delta,\theta) \geq \frac{4Ld(x^*)}{\delta\theta^2} - 1$). Since $k^*_{Switch}(\delta,\theta) \in \left[\frac{4Ld(x^*)}{\delta\theta^2} - 1, \frac{6Ld(x^*)}{\delta\theta^2}\right]$, we conclude that the complexity of the optimal switching policy (i.e. $k^*_{Switch}(\delta,\theta)$) is of the same order $\Theta\left(\frac{LR^2}{\delta\theta^2}\right)$ as the complexity of the general optimal policy (i.e. $k^*(\delta,\theta)$). The restriction to a switching policy costs at most a factor $\frac{3}{2}$ in term of complexity.

3. **When $\theta \geq \theta_{\mathbf{r}}$**:
   Let us come back to the sequence $\{\hat{\alpha}_i\}_{i\geq 0}$, the feasible policy for problem $OptPol(\delta,\theta)$ with the highest coefficients. The definition of $\{\hat{\alpha}_i\}_{i\geq 0}$ is not explicit but we have $\hat{\alpha}_i \simeq \tilde{\alpha}_i = \frac{i+2}{2}$. Therefore, if we add to the general optimization problem $OptPol(\delta,\theta)$, the additional constraints $\alpha_i \leq \frac{i+2}{2}$ for all $i = 0,...,k$, we modify only slightly this problem (we replace the implicit constraint $\alpha_i \leq \hat{\alpha}_i$ by the explicit one $\alpha_i \leq \tilde{\alpha}_i$).

   With this new constraints, we obtain the problem $OptAddPol(\delta,\theta)$:

$$k^*_{add}(\delta,\theta) = \min_{k,\alpha} k \tag{6.31}$$

   such that:

$$\xi(k,\alpha) \geq 0 \tag{6.32}$$

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i = 0,...,k \tag{6.33}$$

$$1 \leq \alpha_i \leq \frac{i+2}{2}, \quad \forall i = 0,...,k. \tag{6.34}$$

   We have that $OptaddPol(\delta,\theta)$ is a restriction of $OptPol(\delta,\theta)$ and a relaxation of $OptSwitchPol(\delta,\theta)$. Therefore $k^*_{Switch}(\delta,\theta) \geq k^*_{add}(\delta,\theta) \geq k^*(\delta,\theta)$. For the problem $OptAddPol(\delta,\theta)$, using the same argument as in Theorem 6.4, we can prove that if there exists $k \geq 0$ such that $\tilde{\alpha}_k \leq \frac{\theta}{2}$ (i.e. $k \leq \theta - 2$) and $\xi(k,\tilde{\alpha}) \geq 0$ then the policy $\tilde{\alpha}$ is optimal.

But we know that if $\theta \geq \theta_r$ then there exists $k \in [0, \theta - 2]$ such that $\xi(k, \tilde{\alpha}) \geq 0$. We conclude that in the case $\theta \geq \theta_r$, the FGM with coefficients $\tilde{\alpha}_i = \frac{i+2}{2}$ for all $i = 0, ..., k$ is optimal, not only for the subset of switching policies (i.e. for the problem $OptSwitchPol(\delta, \theta)$), but also for the more general set of policies that does not grow faster than $\tilde{\alpha}_i$ (i.e. for the problem $OptaddPol(\delta, \theta)$). As the problems $OptaddPol(\delta, \theta)$ and $OptPol(\delta, \theta)$ are almost the same ($\tilde{\alpha}_i$ and $\hat{\alpha}_i$ being of the same order), we conclude that the fast gradient method is (almost) optimal in the case $\theta \geq \theta_r$.

6.4.5 Conclusion: Optimal Switching Policy

Assume that $\theta = \frac{\epsilon}{\delta}$ is an integer such that $1 \leq \theta \leq \frac{Ld(x^*)}{\delta} + 1$ and let $\theta_r = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ be the unique root of $R(\theta) = 2\frac{\theta^3}{3} + \frac{\theta^2}{2} - \frac{13\theta}{6} + 1 - \frac{4Ld(x^*)}{\delta}$. With these assumption, the optimal switching policy (which is also almost optimal without the restriction to switching policies) can be summarized by

- **If $\mathbf{1 \leq \theta \leq 2}$**
  - ◇ Coefficients: $\alpha_i = B_i = 1$, for all $i \geq 0$ (DGM)
  - ◇ Needed number of iterations: $k^*_{Switch}(\delta, \theta) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1$
  - ◇ Optimal Policy ? Yes, $k^*(\delta, \theta) = k^*_{Switch}(\delta, \theta)$.
- **If $\mathbf{2 \leq \theta \leq \theta_r}$**
  - ◇ Coefficients:

$$\alpha_i = \frac{i+2}{2} \text{ for all } i \leq \theta - 2 \text{ and } \alpha_i = \frac{\theta}{2} \text{for all } i \geq \theta - 2$$

$$B_i = \alpha_i^2, \quad \text{for all } i \geq 0.$$

  - ◇ Needed number of iterations:

$$\begin{aligned} k^*_{Switch}(\delta, \theta) &= \frac{4Ld(x^*)}{\delta\theta^2} + \frac{1}{\theta^2}\left(\frac{\theta^3}{3} - \frac{5\theta^2}{2} + \frac{13\theta}{6} - 1\right) \\ &\in \left[\frac{4Ld(x^*)}{\delta\theta^2} - 1, \frac{6Ld(x^*)}{\delta\theta^2}\right]. \end{aligned}$$

  - ◇ Optimal Policy ? Up to a constant factor (at most $\frac{3}{2}$), $k^*(\delta, \theta) \geq \frac{4Ld(x^*)}{\delta\theta^2} - 1$.
- **If $\theta \geq \theta_r$**

◇ Coefficients:

$$\alpha_i = \frac{i+2}{2} \text{ and } B_i = \alpha_i^2 \text{ for all } i \geq 0 \quad (FGM)$$

◇ Needed number of iterations:

$$\begin{aligned} k^*_{Switch}(\delta,\theta) &\in \left[2\sqrt{\frac{Ld(x^*)}{\delta(\theta-1)}} - 4, 2\sqrt{\frac{2Ld(x^*)}{\delta(\theta-2)}}\right] \\ &= \Theta\left(\sqrt{\frac{LR^2}{\theta\delta}}\right). \end{aligned}$$

◇ Optimal Policy ? Almost optimal, $k^*(\delta,\theta) = \Theta\left(\sqrt{\frac{LR^2}{\theta\delta}}\right)$
Optimal if restriction to coefficients such that $\alpha_i \leq \frac{i+2}{2}$.

We see that, using an optimal switching policy, the IGM leads to an improvement compared to the existing methods (i.e. DGM and FGM) when the ratio between target accuracy and oracle accuracy, $\theta = \frac{\epsilon}{\delta}$, is between 2 and θ_r. In addition to its continuity, a remarkable property of the optimal switching policy is the fact that the switching moment has a very simple expression $m = \theta - 2$ which does not depend on L or $d(x^*)$ but only on the ratio between target accuracy and oracle accuracy $\theta = \frac{\epsilon}{\delta}$. For this reason, this method is particularly easy to implement.

6.5 Improvement compared with existing methods

When we are in the range of improvement $[2, \theta_r]$, reaching a target accuracy ϵ can be done in $\Theta\left(\frac{LR^2}{\delta\theta^2}\right) = \Theta\left(\frac{LR^2\delta}{\epsilon^2}\right)$ iterations using the IGM instead of $\Theta\left(\frac{LR^2}{\epsilon-\delta}\right)$ iterations for the GM.
In order to measure the importance of this improvement, we have therefore to answer the two following questions:

1. How extended is the range $[2, \theta_r]$? Is it natural to expect a relative accuracy θ in this interval ?
2. When $\theta \in [2, \theta_r]$, how important is the difference between complexity $\Theta\left(\frac{LR^2}{\delta\theta^2}\right)$ and complexity $\Theta\left(\frac{LR^2}{\epsilon-\delta}\right)$? Is this improvement really significant ?

We would like to point out the fact that the results presented in this section correspond to the worst-case theoretical bounds obtained in the previous section and not to numerical results (whose last section of this chapter will be devoted).

Importance of the range $\theta \in [2, \theta_r]$

The threshold θ_r depends on L, on R and on δ. Let us scale the optimization problem such that $L = 1$ and $R = 1$ and for a given δ, let us define $\epsilon_{MIN} = 2\delta$ and $\epsilon_{MAX} = \theta_r \delta$, respectively the minimal and maximal target accuracies for which IGM leads to an improvement.

For different oracle accuracies, the following table contains the range of improvement of the IGM (with optimal switching policy):

δ	θ_r	ϵ_{MIN}	ϵ_{MAX}
5e-9	1063	1e-8	5.31e-6
5e-8	493	1e-7	2.47e-5
5e-7	228.70	1e-6	1.14e-4
5e-6	106.03	1e-5	5.30 e-4
5e-5	49.10	1e-4	2.5e-3
5e-4	22.69	1e-3	1.13 e-2
5e-3	10.47	1 e-2	5.24e-2
5e-2	4.88	1e-1	2.44e-1
5e-1	2.41	1	1.20

which is represented in a loglog plot in the following figure:

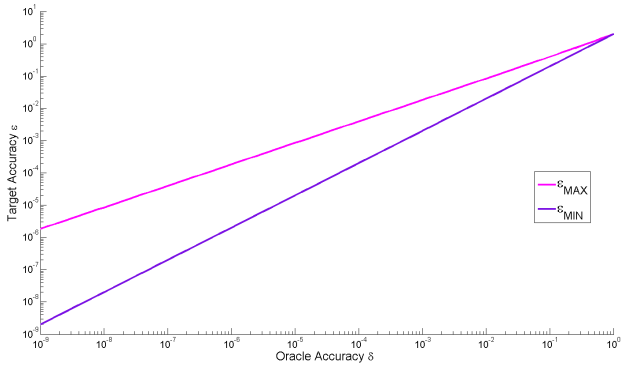


Figure 5: Range of improvement for the IGM with optimal switching strategy.

The size of the interval increases when the oracle accuracy decreases.

If for poor oracle accuracy ($\delta = 5e-3$ or bigger), the range is quite small, we see that for medium oracle accuracy ($\delta = 1e-6$) and high oracle accuracy ($\delta = 1e-9$), the interval $[\epsilon_{MIN}, \epsilon_{MAX}]$ is far from being negligible.

When we are interested in target accuracies that are not too close to the oracle accuracy but at the same time not too poor, we are typically in the range of improvement of the IGM. If we accept to lose one or two digits of accuracy compared to the oracle accuracy, we can take advantage of the new developed IGM in order to reduce the needed number of iterations.

Gain in term of complexity

Let us consider in our discussion four situations: the objective function f is endowed with an exact oracle ($\delta = 0$), an inexact oracle with high accuracy ($\delta = 5e-9$), an inexact oracle with intermediate accuracy ($\delta = 5e-6$) or an inexact oracle with poor accuracy ($\delta = 5e-3$).

For each case, we compare the complexity of the GM, the FGM and the IGM with optimal switching for different target accuracies. More precisely, by GM, we consider in fact the DGM which is nothing else than a IGM but with a switching at the beginning i.e. $\alpha_i = 1$ for all $i \geq 0$. On the other hand, the FGM that we consider here is also a IGM but with $\alpha_i = \frac{i+2}{2}$ for all $i \geq 0$ i.e. without any switching.

1. Exact Oracle $\delta = 0$

When the oracle is exact, the IGM is nothing else that the FGM that clearly outperforms the GM:

ϵ	Complexity GM	Complexity FGM
1e-8	1e8	1.99e4
1e-7	1e7	6.32e3
1e-6	1e6	1.99e3
1e-5	1e5	6.3e2
1e-4	1e4	1.98e2
1e-3	1e3	61
1e-2	1e2	18
1e-1	10	4

The following picture shows us in a loglog plot the clear advantage of the FGM compared to GM in the exact case, when the first-order oracle is exact. When there is no noise, the FGM can reach any accuracy $\epsilon > 0$ and the corresponding needed number of iterations proportional to $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ is highly better than

using the GM (proportional to $\Theta\left(\frac{LR^2}{\epsilon}\right)$)

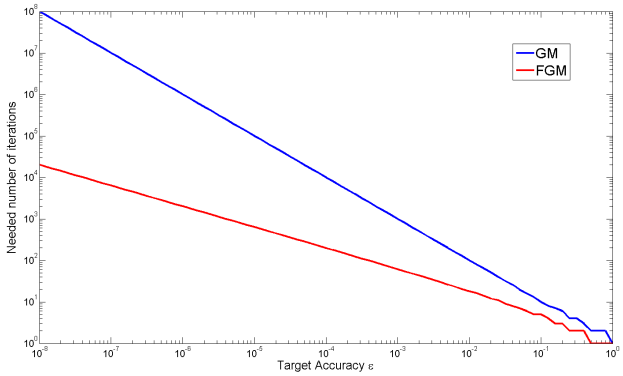


Figure 6: Complexity of the GM and the FGM when used with an exact oracle.

2. Inexact Oracle with high accuracy $\delta = 5e-9$

Let us now assume that the oracle is inexact but with high accuracy, namely $\delta = 5e-9$. In this case, we know that the FGM cannot reach accuracy better that $\epsilon^*_{FGM} = 4.22e-6$ and that the IGM leads to an improvement compared to existing methods in the range $[\epsilon_{MIN}, \epsilon_{MAX}] = [1e-8, 5.312e-6]$. In particular, we point out the fact that the range of improvement of the IGM covers 2.5 orders of magnitude in term of target accuracies.

Remark 6.10. We also see that, whereas these two quantities are of the same order, the best reachable accuracy by the FGM $\epsilon^*_{FGM} = 4.22e-6$ is a little bit smaller than the largest accuracy for which the IGM leads to an improvement $\epsilon_{MAX} = 5.312e-6$. It means that there exists a (small) interval of accuracies reachable by the FGM but for which it is preferable to use an IGM.

Depending on the target accuracy ϵ, the following table contains the corresponding needed number of iterations using the GM, the FGM, the IGM (with optimal switching policy) and the optimal switching moment m for the IGM:

ϵ	Compl. GM	Compl. FGM	Compl. IGM	Switch. Mom. IGM
1e-8	2e8	/	2e8	0
1e-7	1.05e7	/	2e6	18
1e-6	1e6	/	2.01e4	1.98e2
1e-5	1e5	6.69e2	6.69e2	6.69e2
1e-4	1e4	1.98e2	1.98e2	1.98e2
1e-3	1e3	61	61	61
1e-2	1e2	18	18	18
1e-1	10	4	4	4

We can distinguish three different situations:

• When looking for 8 digits of accuracy i.e. to a target accuracy $\epsilon = 1e-8$ (which corresponds exactly to $\epsilon_{MIN} = 2\delta$), we cannot do better than the slow GM. The optimal switching moment is $m = 0$ meaning that we switch to constant coefficients from the beginning and we obtain no improvement compared to the GM. When we are high demanding, looking to a target accuracy close to the oracle accuracy, there is no miracle. Only a very robust and therefore also very slow method like the GM can be used.

• When looking for 7 or 6 digits of accuracy, we are in the range of improvement of the IGM. The FGM cannot be used anymore and the GM is very slow. Compared to the GM, the IGM method allows us to divide by a factor 5 the needed number of iterations in the case $\epsilon = 1e-7$ and even by a factor 50 in the case $\epsilon = 1e-6$. We conclude here that this improvement in term of complexity provided by the new developed IGM is far from being negligible. We observe also that, whereas the optimal switching moment is not anymore zero, it is still small compared to the total needed number of iteration. For 7 digits of accuracy, from the 2 millions of iterations that we have to perform, only the 18th first iterations are 'fast'-type iterations (i.e. with linearly growing coefficients). For 6 digits of accuracy, the proportion increases but remains small, from the 20100 needed iterations, 198 iterations are fast. It is interesting to observe that the division by a factor 50 of the needed number of iterations is obtained using only 1 percent of fast-type iterations at the beginning.

• When looking for 5 or less digits of accuracy, the FGM can reach such level of accuracy with a complexity that cannot be improved using the IGM. In the IGM, the optimal switching moment corresponds exactly to the needed number of iteration, meaning that we never switch to constant coefficients and the IGM with optimal switching policy is nothing else that the FGM. When we are happy with a not so accurate solution, the FGM allows us to solve the problem with an unbeatable complexity proportional to $O\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$.

A loglog plot of the respective complexities of GM, FGM and IGM is perhaps the best way to illustrate the improvement obtained using the IGM (with optimal switching policy):

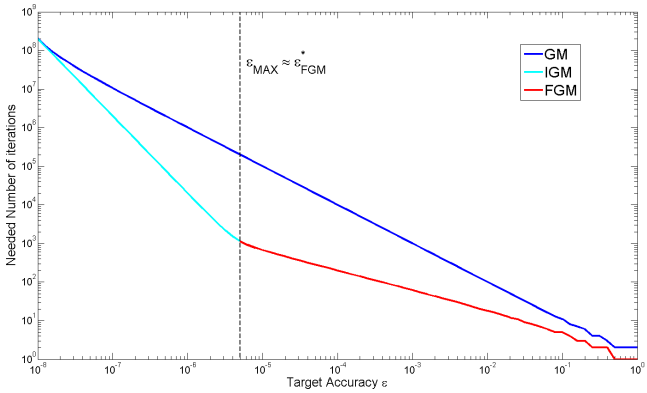


Figure 7: Complexities of the GM, the FGM and the IGM with oracle accuracy $\delta = 5e-9$.

The IGM allows us to obtain target accuracy unreachable by the FGM in a significantly smaller amount of iterations compared to the GM.

On the range of improvement of the IGM, the gain compared to the GM and the proportion of fast steps used in the IGM increases when the target accuracy increases (i.e. becomes less good). When the target accuracy becomes close to ϵ_{MAX}, the complexity of the IGM becomes similar to the complexity of the FGM in the exact case and the proportion of fast-type iterations tends to 1. The followings loglog plots illustrates these phenomena:

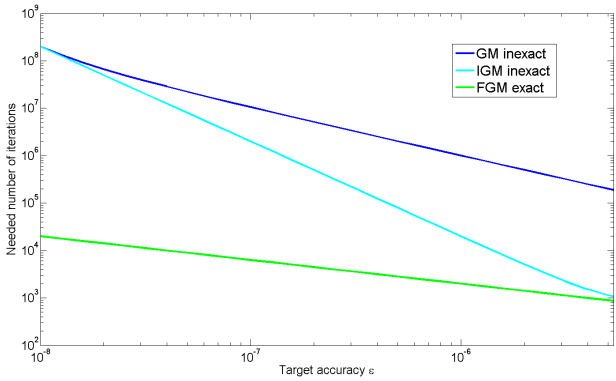


Figure 8: $\delta = 5e-9$: Complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement of the IGM.

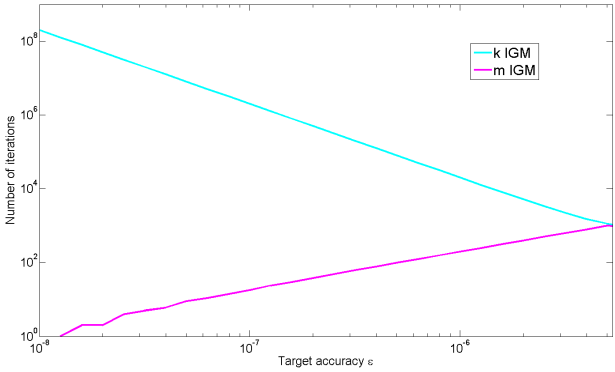


Figure 9: $\delta = 5e-9$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

3. Inexact Oracle with medium accuracy $\delta = 5e-6$

In this case, we have $\epsilon^*_{FGM} = 4.22e-4$, $\epsilon_{MIN} = 1e-5$ and $\epsilon_{MAX} = 5.30e-4$. The range of improvement of the IGM $[1e-5, 5.30e-4]$ is reduced, covering 1.5 orders of magnitude. In the remainder, the same kind of behavior than with the accurate oracle is obtained, as proved by the following table and plots:

ϵ	Compl. GM	Compl. FGM	Compl. IGM	Switch. Mom. IGM
1e-5	2e5	/	2e5	0
1e-4	1.05e4	/	2e3	18
1e-3	1e3	65	65	65
1e-2	1e2	18	18	18
1e-1	10	4	4	4

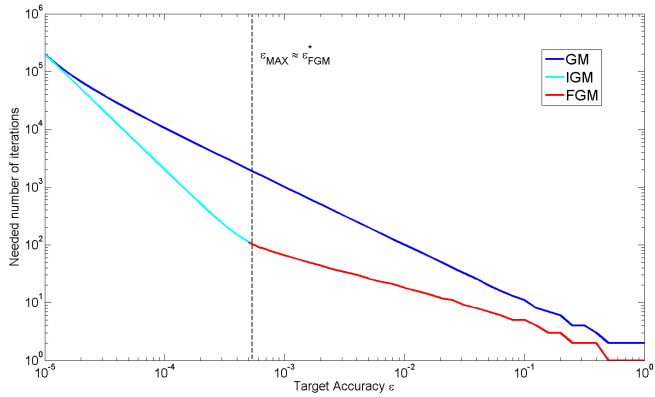


Figure 10: Complexities of the GM, the FGM and the IGM with oracle accuracy $\delta = 5e-6$.

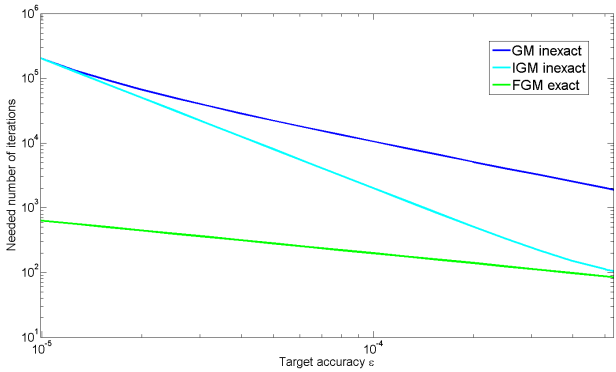


Figure 11: $\delta = 5e-6$: Complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement of the IGM.

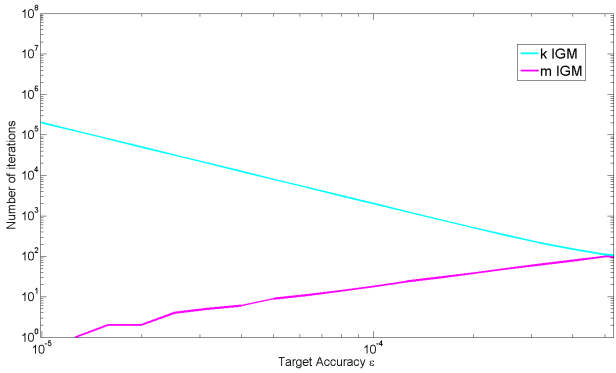


Figure 12: $\delta = 5e-6$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

4. Inexact Oracle with poor accuracy $\delta = 5e-3$

In this case, we have $\epsilon^*_{FGM} = 4.25e-2$, $\epsilon_{MIN} = 1e-2$ and $\epsilon_{MAX} = 5.24e-2$. The range of improvement of the IGM $[1e-2, 5.24e-2]$ is again reduced, covering now only 0.5 orders of magnitude. When the target accuracy lies in the range of improvement, we retrieve a similar behavior than with the other levels of oracle accuracy:

ϵ	Compl. GM	Compl. FGM	Complexity IGM	Switch. Mom. IGM
1e-2	2e2	/	200	0
1e-1	11	5	5	5

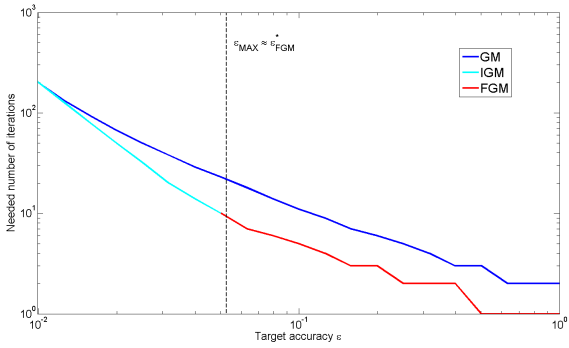


Figure 13: Complexities of the GM, the FGM and the IGM with oracle accuracy $\delta = 5e-3$.

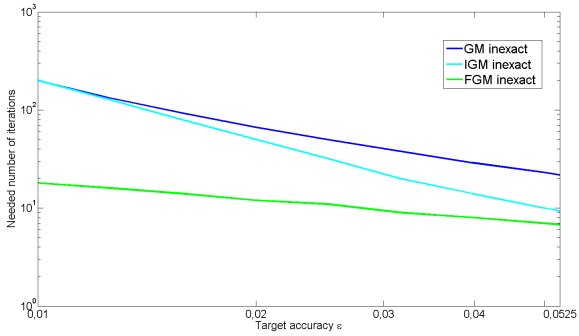


Figure 14: $\delta = 5e-3$: Complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement.

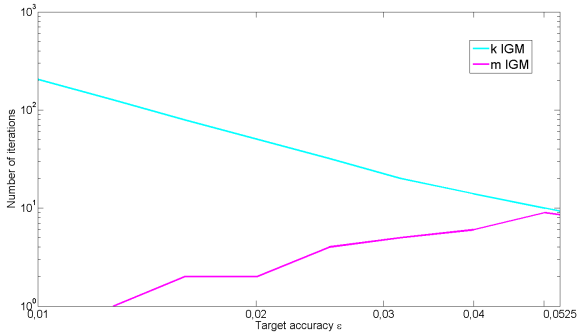


Figure 15: $\delta = 5e-3$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

6.6 Choice of the coefficients for an intermediate convergence rate

6.6.1 Power Policy

In this section, our goal is to obtain, with a good choice for the sequences of coefficients $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$, a family of methods exhibiting the whole spectrum of convergence rates given by Theorem 4.12. More precisely, for all $p \in [1,2]$, we want to obtain a method with intermediate convergence rate in the exact case $\Theta(\frac{LR^2}{k^p})$ and corresponding optimal rate of errors accumulation $\Theta(k^{p-1}\delta)$. We know that the general convergence rate of the IGM is given by:

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{\sum_{i=0}^k \alpha_i} + \frac{\sum_{i=0}^k B_i}{\sum_{i=0}^k \alpha_i}\delta. \tag{6.35}$$

Therefore, we would like to find a feasible sequence of coefficients $\{\alpha_i\}_{i\geq 0}$ such that:

1.
$$\sum_{i=0}^k \alpha_i = \Theta(k^p) \tag{6.36}$$

and

2.
$$\frac{\sum_{i=0}^k B_i}{\sum_{i=0}^k \alpha_i} = \Theta(k^{p-1}) \tag{6.37}$$

The condition (6.36) suggests to choose $\alpha_i = \Theta(i^{p-1})$. If we are able to obtain a feasible sequence $\alpha_i = \Theta(i^{p-1})$ such that $\alpha_i^2 \leq A_i$ and $\alpha_i \geq 1$ then we could take $B_i = \alpha_i^2 = \Theta(i^{2p-2})$. With this choice, we would have $\sum_{i=0}^k B_i = \Theta\left(k^{2p-1}\right)$ and therefore $\frac{\sum_{i=0}^k B_i}{\sum_{i=0}^k \alpha_i} = \Theta(k^{p-1})$.

This reasoning leads us to choose $\alpha_i = \left(\frac{i+p}{p}\right)^{p-1}$ and $B_i = \alpha_i^2$ for all $i \geq 0$. With this choice, we have: $\sum_{i=0}^k \alpha_i \geq \int_0^k \left(\frac{x+p}{p}\right)^{p-1} dx + \alpha_0 = \left(\frac{k+p}{p}\right)^p$. Therefore our sequence $\{\alpha_i\}_{i\geq 0}$ is feasible:

$$\alpha_k^2 = \left(\frac{k+p}{p}\right)^{2p-2} = B_k \leq \left(\frac{k+p}{p}\right)^p \leq A_k = \sum_{i=0}^k \alpha_i, \quad \forall k \geq 0$$

and

$$B_k \geq \alpha_k, \quad \forall k \geq 0.$$

Furthermore, we have:

$$\begin{aligned}\sum_{i=0}^{k} B_i = \sum_{i=0}^{k} \alpha_i^2 &\leq \int_0^k \left(\frac{x+p}{p}\right)^{2p-2} dx + \left(\frac{k+p}{p}\right)^{2p-2} \\ &\leq \frac{p}{2p-1}\left(\frac{k+p}{p}\right)^{2p-1} + \left(\frac{k+p}{p}\right)^{2p-2} \\ &\leq \left(\frac{k+p}{p}\right)^{2p-1} + \left(\frac{k+p}{p}\right)^{2p-2}.\end{aligned}$$

We conclude that with this choice of coefficients, the IGM exhibits the wanted convergence rate:

$$\begin{aligned}f(y_k) - f^* &\leq Ld(x^*)\left(\frac{p}{k+p}\right)^p + \left(\left(\frac{k+p}{p}\right)^{p-1} + \left(\frac{k+p}{p}\right)^{p-2}\right)\delta \\ &\leq Ld(x^*)\left(\frac{p}{k+p}\right)^p + \left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta \\ &= \Theta\left(\frac{Ld(x^*)}{k^p}\right) + \Theta(k^{p-1}\delta).\end{aligned}$$

Some of these intermediate convergence rates are represented in the following picture:

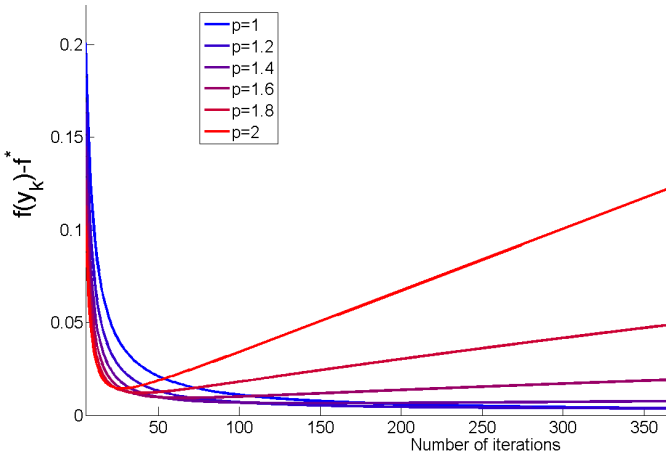


Figure 16: IGM with Power Policy, different convergence rates depending on the choice for p.

For a given $1 \leq p \leq 2$, the method reach its minimum after

$$k_p^* = p\left(\left(\frac{p}{p-1}\right)^{\frac{1}{2p-1}}\left(\frac{Ld(x^*)}{\delta}\right)^{\frac{1}{2p-1}} - 1\right) = \Theta\left(\left(\frac{LR^2}{\delta}\right)^{\frac{1}{2p-1}}\right)$$

iterations.

When $p = 1$, we obtain $k_1^* = +\infty$ since the method is decreasing and therefore reaches its minimal value at the limit. When $p > 1$, the method is decreasing at first, until it reaches its minimum after $k_p^* = \Theta\left(\left(\frac{LR^2}{\delta}\right)^{\frac{1}{2p-1}}\right)$ iterations and increasing after.

The corresponding best reachable accuracy is given by:

$$\begin{aligned}\epsilon_p^* &= \left(\left(\frac{p-1}{p}\right)^{\frac{p}{2p-1}} + \left(\frac{p}{p-1}\right)^{\frac{p}{2p-1}}\right)(Ld(x^*))^{\frac{p-1}{2p-1}}\delta^{\frac{p}{2p-1}} + \delta \\ &= \Theta\left((LR^2)^{\frac{p-1}{2p-1}}\delta^{\frac{p}{2p-1}}\right).\end{aligned}$$

When $p = 1$, we retrieve the Dual Gradient Method (i.e. $\alpha_i = B_i = 1$ for all $i \geq 0$) with convergence rate $\Theta\left(\frac{LR^2}{k}\right)$ in the exact case, no accumulation of error and best reachable accuracy $\epsilon_1^* = \delta$.

When $p = 2$, we retrieve a FGM: optimal convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$ in the exact case, accumulation of error with a rate $\Theta(k\delta)$ and minimum reachable accuracy $\epsilon_2^* = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$ after $\Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$.

For $1 < p < 2$, we obtain new methods with intermediate convergence rate, intermediate rate of error accumulation and intermediate best reachable accuracy.

6.6.2 Optimal power choice

In order to obtain a family of methods with intermediate convergence rate, we have considered a power policy $\alpha_i = \Theta(i^{p-1})$ in the IGM.

Our goal here is to see how a power policy can be also used in practice in order to reach a target accuracy ϵ and to compare its efficiency with the optimal switching policy described in the previous section. For the simplicity of

our further analysis, we use a weaker upper bound for the convergence with coefficients independent of p:

$$\begin{aligned} f(y_k) - f^* &\leq Ld(x^*)\left(\frac{p}{k+p}\right)^p + \left(\frac{k+p}{p}\right)^{p-1}\delta + \delta \\ &\leq Ld(x^*)\left(\frac{2}{k+1}\right)^p + (k+2)^{p-1}\delta + \delta \\ &\leq Ld(x^*)\frac{2^p}{(k+1)^p} + 2^{p-1}(k+1)^{p-1}\delta + \delta \\ &\leq \frac{4Ld(x^*)}{(k+1)^p} + 2(k+1)^{p-1}\delta + \delta = Acc(k,p,\delta). \end{aligned}$$

Remark 6.11. The fact that we need to weaken the upper bound on the convergence in order to be able to analyze the method is one of the drawbacks of the power policy compared to the switching policy.

δ and k fixed

First, we assume that the number of iterations k and the oracle accuracy δ are fixed.
In this case, we can minimize $Acc(k,p,\delta)$ with respect to $p \in [1,2]$. The unconstrained problem $\min_p Acc(k,p,\delta)$ has an optimal solution p^* such that $(k+1)^{2p^*-1} = \left(\frac{2Ld(x^*)}{\delta}\right)$ and therefore $p^* = \frac{1}{2}\left[\frac{\ln\left(\frac{2Ld(x^*)}{\delta}\right)}{\ln(k+1)} + 1\right]$.

However, we need also to satisfy $1 \leq p \leq 2$:

- $p \geq 1$ gives the condition $\ln\left(\frac{2Ld(x^*)}{\delta}\right) \geq \ln(k+1) \Leftrightarrow k \leq \frac{2Ld(x^*)}{\delta} - 1$.
  If $k \geq \frac{2Ld(x^*)}{\delta} - 1$, we take $p = 1$.
- $p \leq 2$ gives the condition $\ln\left(\frac{2Ld(x^*)}{\delta}\right) \leq 3\ln(k+1) \Leftrightarrow k \geq \sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1$.
  If $k \leq \sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1$, we take $p = 2$.

In conclusion, for fixed k and δ, the optimal choice for p is given by:

- $$p(k,\delta) = 2, \quad \text{i.e. a Fast Gradient Method}$$
  if $0 \leq k \leq k_1 = \sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1$
- $$p(k,\delta) = \frac{1}{2}\left[\frac{\ln\left(\frac{2Ld(x^*)}{\delta}\right)}{\ln(k+1)} + 1\right]$$

if $\sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1 = k_1 \leq k \leq k_2 = \frac{2Ld(x^*)}{\delta} - 1$

- $$p(k, \delta) = 1, \quad \text{i.e. the Dual Gradient Method}$$

  if $k \geq k_2 = \frac{2Ld(x^*)}{\delta} - 1$.

The accuracy on the objective function that we obtain with this optimal choice for p is therefore:

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)^2} + 2(k+1)\delta + \delta := Acc(k, p(k, \delta), \delta)$$

when $0 \leq k \leq k_1 = \sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1$

$$f(y_k) - f^* = \frac{4\sqrt{2}\sqrt{Ld(x^*)\delta}}{\sqrt{k+1}} + \delta := Acc(k, p(k, \delta), \delta)$$

when $\sqrt[3]{\frac{2Ld(x^*)}{\delta}} - 1 = k_1 \leq k \leq k_2 = \frac{2Ld(x^*)}{\delta} - 1$

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{k+1} + 3\delta := Acc(k, p(k, \delta), \delta)$$

when $k \geq k_2 = \frac{2Ld(x^*)}{\delta} - 1$.

The function $BestAcc(k, \delta) = Acc(k, p(k, \delta), \delta)$ is continuous in k. Indeed, we have $BestAcc(k_1) = 2^{7/3}(Ld(x^*))^{1/3}\delta^{2/3} + \delta$ and $BestAcc(k_2) = 5\delta$. Furthermore, this function is clearly decreasing in k on the intervals $[k_1, k_2]$ and $[k_2, +\infty[$. On the interval $[0, k_1]$,$BestAcc(k) = \frac{4Ld(x^*)}{(k+1)^2} + 2(k+1)\delta + \delta$. This function is convex and reach its unique minimum at the point $k^* = 2\sqrt[3]{\frac{Ld(x^*)}{\delta}} > k_1$. Therefore $BestAcc(.)$ is a decreasing function of k on $[0, +\infty[$.

δ and ϵ fixed

We assume now that the oracle accuracy δ and the needed accuracy for the objective function ϵ are fixed whereas the number of iteration k and the parameter p can be chosen. We want, by a good choice of p, to minimize the number of iteration k needed to reach an accuracy ϵ:

$$\min_{k \geq 0, p \in [1,2]} k, \text{ s.t. } Acc(k, p, \delta) \leq \epsilon$$

or equivalently:

$$\min_{k \geq 0} k, \text{ s.t. } BestAcc(k, \delta) \leq \epsilon.$$

We conclude that, if we want an accuracy for the objective function of ϵ (i.e $f(y_k) - f^* \leq \epsilon$) such that:

1. $\epsilon \geq BestAcc(k_1) = 2^{7/3}(Ld(x^*))^{1/3}\delta^{2/3} + \delta$, we have to choose $p = 2$ i.e. the Fast Gradient Method (FGM).
2. $BestAcc(k_2) = 5\delta \leq \epsilon \leq BestAcc(k_1) = 2^{7/3}(Ld(x^*))^{1/3}\delta^{2/3} + \delta$, we have to use an intermediate value of $p \in]1, 2[$. The needed number of iterations is $k(\epsilon, \delta) = \frac{32Ld(x^*)}{(\epsilon-\delta)^2} - 1$ and the parameter of the method is: $p(\epsilon, \delta) = \frac{1}{2}\left[\frac{ln\left(\frac{2Ld(x^*)}{\delta}\right)}{\ln\left(\frac{32Ld(x^*)}{(\epsilon-\delta)^2}\right)} + 1\right]$.
3. $\delta \leq \epsilon \leq BestAcc(k_2) = 5\delta$, we have to choose $p = 1$ i.e. the Dual Gradient Method (DGM).

Like with the switching policy, we obtain three regimes depending on the ration between ϵ and δ. When $\epsilon \leq \epsilon_1 = \Theta(\delta)$, we have to use the Dual Gradient Method with complexity $\Theta\left(\frac{LR^2}{\epsilon}\right)$. When $\epsilon \geq \epsilon_2 = \Theta((LR^2)^{1/3}\delta^{2/3})$, we have to use a Fast Gradient Method with complexity $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$. For intermediate target accuracy $\epsilon_1 \leq \epsilon \leq \epsilon_2$, we have to use a method with intermediate behavior and with complexity $\Theta\left(\frac{LR^2\delta}{\epsilon^2}\right)$.

However, compared to the switching policy, the absolute constant factor in the complexity of the power policy is less favorable (this difference is perhaps partially due to an analysis based on a weaker upper bound). Furthermore, in the intermediate regime, the optimal choice of p depends on L and R, which is not the case of the optimal switching moment $m = \theta - 2$. The optimal switching policy is typically easier to implement than the optimal power policy.

6.7 Numerical Illustration

Let us finish this chapter with a small numerical experiment. Our goal is to observe on a practical example the main results obtained in this chapter (and in Chapter 4). We consider the situation of a convex quadratic function on the unit simplex

$$\min_{x \in \Delta_n} \frac{1}{2} x^T A x \tag{6.38}$$

where $\Delta_n = \{x \in \mathbb{R}^n_+ | \sum_{i=1}^n x^{(i)} = 1\}$ and $n = 1000$. The matrix A is chosen such that its minimal eigenvalue $\lambda_{min}(A) = 0$ in order to avoid any strong convexity property.

To solve this problem, we use the intermediate gradient method, more precisely its variant with only prox-type subproblems, that we have developed in subsection 6.1.2. We choose the l_1 setup i.e. we work with the l_1 norm $\|.\|_E = \|.\|_1$

and the entropy prox-function $d(x) = \ln(n) + \sum_{i=1}^{n} x^{(i)} \ln(x^{(i)})$. We perform a fixed number of iterations $k = 500$ and consider different choices for the sequence of coefficients $\{\alpha_i\}_{i \geq 0}$:

- ◇ Constant stepsize $\alpha_i = 1$ for all $i \geq 0$ for which this method is nothing else than the non-Euclidean DGM developed in [25] and described in subsection 2.4.4. In this section, we use the generic name GM for this method.

- ◇ Linearly growing coefficients $\alpha_i = \frac{i+2}{2}$ for all $i \geq 0$ for which this method is nothing else than a FGM with Bregman distance comparable with the method developed in [59] and described in subsection 2.4.5.

- ◇ Switching coefficients

$$\alpha_i = \begin{cases} \frac{i+2}{2} & \text{when } i = 0, ..., m \ , \\ \frac{m+2}{2} & \text{when } i = m+1, ..., k \end{cases}$$

  with $m = 5, 50$ or 250 corresponding respectively to $1\%, 10\%$ and 50% of fast-type iterations. (With 0% of fast-type iterations, we retrieve the GM and with 100%, the FGM.)

- ◇ Power coefficients $\alpha_i = \left(\frac{i+p}{p}\right)^{p-1}$ with $p = 1.2, 1.4, 1.6$ and 1.8. (With $p = 1$, we retrieve the GM and with $p = 2$, the FGM.)

We compare these methods on the problem (6.38) when used with an approximate gradient $g_{\delta,L}(y) = Ay + \xi$ with $\|\xi\|_\infty = \frac{\delta}{2\text{diam}\,(\Lambda_n)} = \frac{\delta}{4}$ and $L = \lambda_{\max}(A) = 1$ (this kind of approximate gradient leads to a (δ, L)-oracle as we have seen in subsection 4.1.3). Three levels of errors are considered: $\delta = 0$, $\delta = 1e-2$ and $\delta = 1e-1$.

6.7.1 Behavior with exact oracle $\delta = 0$

With an exact oracle, the FGM is unbeatable, the GM is significantly slower and the intermediate methods exhibit intermediate behaviors as announced by the theory. When the switching policy is used, fastness of the method increases with the number of fast-type iterations m. With the power policy, fastness increases when p increases between 1 and 2.

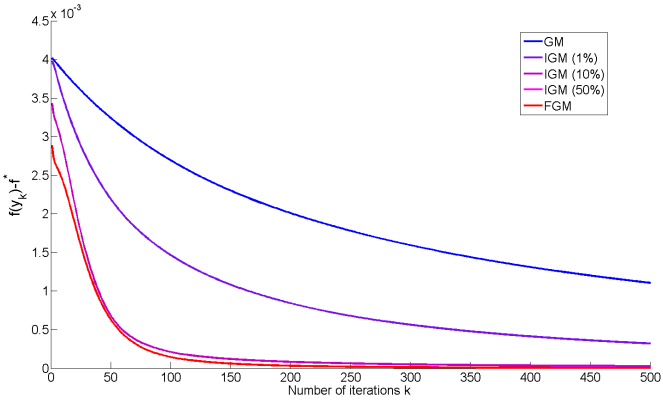


Figure 17: IGM with switching policy: behavior in the exact case

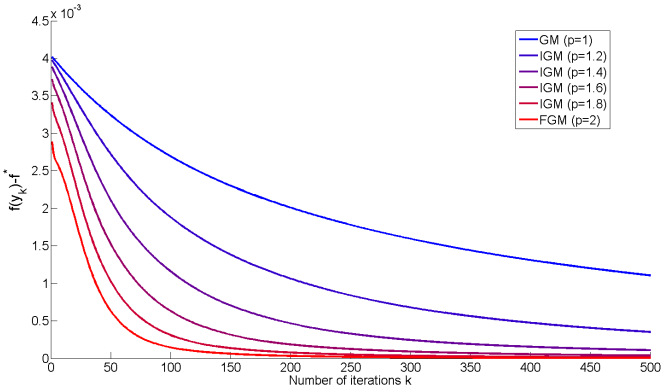


Figure 18: IGM with power policy: behavior in the exact case

Remark 6.12. Even if all these methods start with the same initial point x_0, these plots have different value at $k = 0$. This comes from the fact that y_0 is already an iterate generated by the methods and differs therefore from one method to another one.

6.7.2 Introducing errors in the first-order methods: behavior with $\delta = 1e - 2$

When errors are introduced in the first-order information, we observe the behavior described by theory:

- ◇ A GM which is slow but robust with respect to errors

- ◇ A FGM which suffers from an higher sensitivity with respect to oracle errors
- ◇ The Intermediate gradient methods that exhibit intermediate fastness and intermediate robustness with respect to errors. When m or p increases, the IGM becomes faster at the beginning but the effect of the accumulation of errors becomes also more quickly dominant.
- ◇ With a well chosen value of m in the switching policy or of p in the power policy, the corresponding IGM outperforms the GM and FGM, it can reach accuracies unreachable by the FGM in a significantly smaller amount of time compared with the GM.

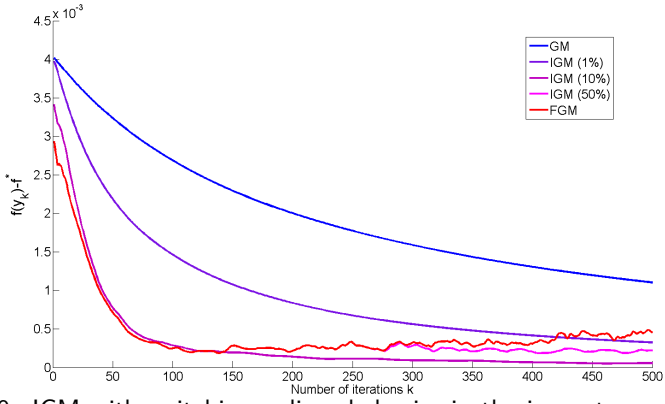


Figure 19: IGM with switching policy: behavior in the inexact case with $\delta = 1e-2$

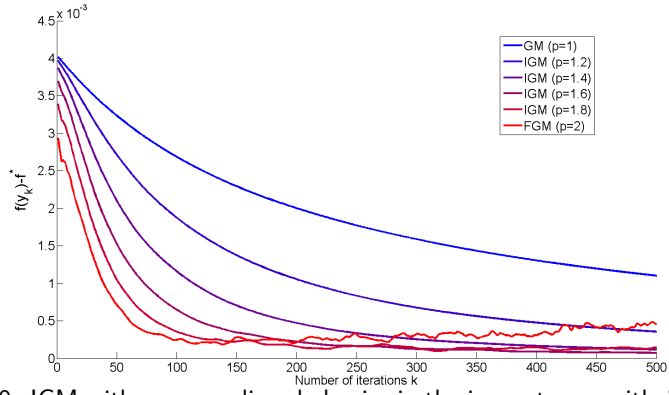


Figure 20: IGM with power policy: behavior in the inexact case with $\delta = 1e-2$

6.7.3 Increasing the oracle errors: behavior with $\delta = 1e-1$

When we increase the level of the oracle errors, the effect of the errors in the convergence rate logically increases and we have, as predicted by the theory, to reduce the number of fast-type iterations in the switching policy and the value of p in the power policy:

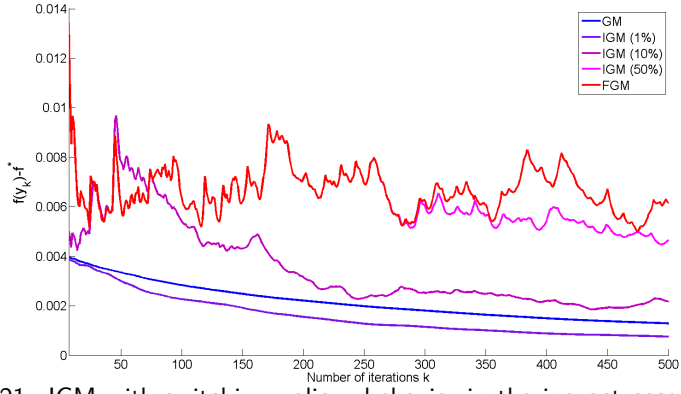


Figure 21: IGM with switching policy: behavior in the inexact case with $\delta = 1e-1$

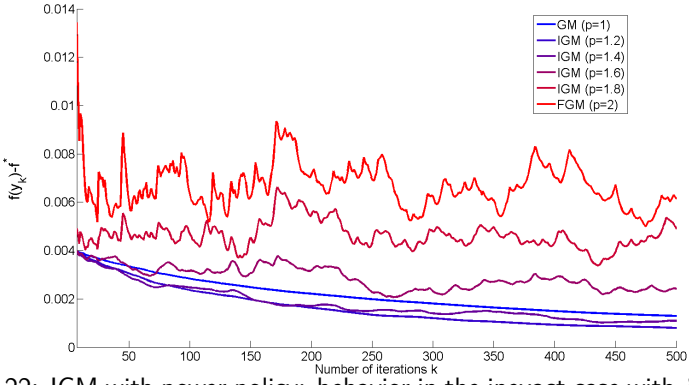


Figure 22: IGM with power policy: behavior in the inexact case with $\delta = 1e-1$

Chapter 7

Stochastic First Order Methods in Smooth Convex Optimization

This chapter corresponds to the paper [25]:
O. Devolder. **Stochastic First Order Methods in Smooth Convex Optimization** *CORE Discussion Paper 2011/70*, (2011).

Chapter 7 in two Questions/Answers

◇ *Is the case of a stochastic inexact oracle in some sense more favorable compared with the deterministic one ?*

Yes and No. If we do not want to modify the existing first-order methods, more precisely if we continue to use them with constant stepsizes, the behaviors of the PGM/DGM/FGM in the stochastic case are similar to what we have obtained in the deterministic case. The stochastic noise (multiplied by the diameter of the feasible set) simply replaces the deterministic error and the expected convergence rate replaces the deterministic one.

However, the stochastic case is in some sense more favorable than the deterministic one (at least when the oracle is unbiased, otherwise the bias playing the role of a deterministic error). Taking into account the stochastic nature of the first-order information in the first-order methods, it is possible to improve the performance of these methods.

◇ *Is it possible to modify the existing first-order methods of smooth convex optimization in order to reduce the effect of a stochastic noise to zero ? At which rate ?*

Yes! Whereas we cannot expect to obtain an accuracy on the objective

function better than δ in the deterministic case, it is possible to expect a decrease of the stochastic noise effect up to zero with a rate proportional to $\frac{1}{\sqrt{k}}$. Furthermore, this rate of decrease of the stochastic noise effect is unimprovable and can be obtained simply by using well-chosen decreasing stepsizes in the PGM/DGM/FGM (that become with these modifications respectively the SPGM, the SDGM and the SFGM). In particular, the faster convergence rate of the SFGM does not prevent us to decrease the stochastic noise at the same rate as for the SPGM/SDGM.

Contents

7.1 Smooth convex problem with stochastic oracle **241**
7.1.1 Problem class and biased stochastic oracle . . . 241
7.1.2 Examples 243
7.2 Stochastic Primal Gradient Method **245**
7.2.1 Scheme 245
7.2.2 General Convergence Rate 247
7.2.3 Choice of Stepsizes 250
7.3 Stochastic estimate functions **253**
7.4 Stochastic Dual Gradient Method **254**
7.4.1 Scheme 254
7.4.2 General Convergence rate 256
7.4.3 Choice of the Coefficients 259
7.5 Stochastic Fast Gradient Method **261**
7.5.1 Scheme 261
7.5.2 General convergence rate 263
7.5.3 Choice of the Coefficients 267
7.6 Probability of large deviation **270**
7.6.1 Probability of large deviation for SDGM 272
7.6.2 Probability of large deviation for the SFGM . . 273
7.7 Postoptimization: Accuracy certificate **275**
7.8 Numerical Experiments **278**

This chapter is devoted to the development of efficient first-order methods for convex optimization problems of the form $\min_{x \in Q} f(x)$ where f is a smooth convex function but now endowed with a stochastic first-order oracle.

In the deterministic convex case, smoothness is a highly desirable property. Indeed, for a nonsmooth Lipschitz-continuous function (with constant M), the best convergence rate for $f(y_k) - f^*$ (where k is the iteration counter and y_k the approximate solution generated after k iterations) that we can expect, using

a first-order method, has the form $O\left(\frac{MR}{\sqrt{k}}\right)$ where R represents the distance between the initial iterate and the optimal solution. This slow rate is achieved for example by subgradient type methods (see for example [58, 36]).

On the other hand, when the objective function is smooth with a Lipschitz-continuous gradient (with constant L), the convergence rate of the (sub)gradient method becomes $O\left(\frac{LR^2}{k}\right)$ and it is even possible to obtain a convergence rate $O\left(\frac{LR^2}{k^2}\right)$ (optimal for deterministic smooth problem) using the fast gradient methods developed in various variants by Nesterov and others since 1983 ([56, 57, 58, 59]).

In the stochastic convex case, when the first-order information is affected by a random noise, the most classical first-order methods are the Stochastic Approximation (SA) methods that mimic the subgradient method, replacing the exact gradient by the stochastic one. In the modern SA methods, like the Mirror Descent SA method (see [52]), the function endowed with a stochastic oracle is typically assumed to be nonsmooth and the obtained convergence rate is $O\left(\frac{MR}{\sqrt{k}}+\frac{\sigma R}{\sqrt{k}}\right)$ where σ is the level of the stochastic noise. This rate has an optimal dependence in M (since the problem is nonsmooth) but also an optimal dependence in σ. Indeed, it has been proved in [55] that the effect of the stochastic noise cannot be decreased, by a first-order method, with a better rate than $\frac{1}{\sqrt{k}}$, and this limitation is also valid when the function is smooth. This result has led to the common belief that in the presence of a stochastic oracle, the smoothness of the objective function is useless. It does not matter that the function is smooth or not, in any case we come back to a slow convergence rate $O\left(\frac{1}{\sqrt{k}}\right)$ like in the deterministic nonsmooth case. However when the Lipschitz constant of the gradient L is big as compared to the stochastic noise σ, and when we are interested in solutions with moderate accuracy, a convergence rate of the form $O\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ or $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$, exploiting the smoothness of f in its first term, can be significantly better than $O\left(\frac{MR}{\sqrt{k}}+\frac{\sigma R}{\sqrt{k}}\right)$.

First-order methods in the stochastic smooth case has been considered for the first time by Lan in [44]. In this paper, he adapts the Mirror descent SA method, designed initially for nonsmooth problems, to the smooth case, obtaining a convergence rate of the form $O\left(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}\right)$ and adapts one variant of the fast gradient methods, initially designed for deterministic smooth problems, to the stochastic case, obtaining a convergence rate of the form $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$. The development of fast gradient methods in the smooth stochastic case with ap-

plications in machine learning problems has also been considered recently in [30, 67].

The new first-order methods that we develop in this chapter exhibit the same type of convergence rates but are characterized by some common properties that extend their applicability in practice:

1. Our methods can be used with a general norm (not especially the Euclidean norm) and can be adapted to the problem geometry, using a good setup and therefore easy to solve auxiliary subproblems. This desirable property is not satisfied by the methods developed in [30, 67]. In these papers, the methods use auxiliary subproblems based on the squared norm featuring quadratic functions that are sometimes difficult to minimize on the feasible set.

2. Our methods use stepsizes that do not require to know an a priori knowledge of the performed number of iterations. This property is highly desirable when we want to run a method for a given time and not for a given number of iterations (for example when we compare methods with different iteration costs). On the contrary, the methods developed in [44, 67] assume an a priori knowledge of the performed number of iterations N and use stepsizes based on this number. We discuss in details the stepsize choice for each method in Sections 7.2, 7.4, 7.5 and show in Section 7.8 with numerical experiments, the advantage of stepsizes policies not based on N.

3. Our methods can be applied, without modification of the convergence rate, to the composite case where we add to the smooth objective function f, an easy convex function h (potentially nonsmooth) that can be kept in the auxiliary subproblems used by the first-order methods. This composite case, when f is endowed with a stochastic oracle and h is easy, has been already considered in [67] but not in [44, 30].

   Remark 7.1. Lan in [44] considers a different composite case where the nonsmooth part of the function is also endowed with a stochastic black-box oracle. In our case, we use the explicit structure of the (possibly) nonsmooth component, avoiding in this way that h slows down the convergence rate.

Furthermore, to the best of our knowledge, this chapter considers for the first time, the biased case i.e. the situation where the smooth function f is endowed with an oracle which is not only stochastic (with stochastic noise σ) but also biased (with bias δ), meaning that on average, the stochastic first-order information does not coincide with the exact one.

The chapter is organized as follows. In Section 7.1, we present in a more formal form our problem class and three simple examples of smooth convex problems with stochastic oracle. In some cases, the stochasticity is in the problem since the beginning. In other cases, we introduce ourselves the stochasticity via a randomization of the first-order information in order to reduce the computational cost of the first-order methods. In Section 7.2, we develop new practical stepsize policy for the Stochastic Primal Gradient Method (SPGM) which is nothing else than the Mirror-Descent SA method (see [52, 44]) but applied to a smooth convex problem. In Section 7.3, we generalize the machinery of estimate functions to the stochastic case. Based on this principle, we develop and study the average behavior of two new methods, a Stochastic Dual Gradient Method (SDGM) with convergence rate $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ (Section 7.4) and a Stochastic Fast Gradient Method (SFGM) with convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ (Section 7.5). All these methods decrease the effect of the stochastic noise at the unimprovable rate $O\left(\frac{\sigma R}{\sqrt{k}}\right)$ where R represents the distance between the initial iterate x_0 and the optimal solution set and k is the iteration counter. In Section 7.6 and 7.7, we study the probabilities of large deviations for these methods and develop accuracy certificates. The last section (Section 7.8) is devoted to numerical experiments. We consider quadratic problems on the simplex when the gradient is affected by a stochastic noise and compare our methods (using different possible stepsize policies) with existing methods.

7.1 Smooth convex problem with stochastic oracle

7.1.1 Problem class and biased stochastic oracle

We consider the convex optimization problem:

$$\phi^* = \min_{x \in Q} \phi(x) \tag{7.1}$$

where $Q \subset E$ is a closed convex set, $\phi = f + h$ and

- $\diamond$ $f : Q \to \mathbb{R}$ is a convex function, typically smooth but endowed with a stochastic first-order oracle (possibly biased)
- $\diamond$ $h : Q \to \mathbb{R}$ is an easy convex function. For a well-chosen prox-function $d(.)$, easy means that we can easily solve subproblems of the form

  $$\min_{x \in Q} \{\langle g, x\rangle + Cd(x) + h(x)\}$$

  for all $C > 0$ and $g \in E^*$.

The stochastic first-order oracle available for f is characterized by two levels of inexactness:

- ◇ The function f is endowed with a (δ, L)-oracle i.e. that for each $x \in Q$, we could potentially compute $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$ such that

$$0 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x \rangle \leq \frac{L}{2} \|x - y\|_E^2 + \delta, \quad \forall y \in Q. \tag{7.2}$$

- ◇ We do not use $(f_{\delta,L}(x), g_{\delta,L}(x))$ but instead stochastic estimates $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$. More precisely, at all point $x \in Q$, we associate with x a random variable X whose probability distribution is supported on $\Xi \subset \mathbb{R}^d$ and such that

$$E_{\xi \sim X}[F_{\delta,L}(x,\xi)] = f_{\delta,L}(x) \tag{7.3}$$

$$E_{\xi \sim X}[G_{\delta,L}(x,\xi)] = g_{\delta,L}(x) \tag{7.4}$$

$$E_{\xi \sim X}[(\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_E^*)^2] \leq \sigma^2. \tag{7.5}$$

When $\delta = 0$, f is necessarily smooth with a Lipschitz-continuous gradient (with constant L i.e. $f \in F_L^{1,1}(Q)$) and the oracle is stochastic but unbiased: $E_{\xi \sim X}[F_{\delta,L}(x,\xi)] = f(x)$ and $E_{\xi \sim X}[G_{\delta,L}(x,\xi)] = \nabla f(x)$.

When $\delta \neq 0$, this kind of oracle can be seen as a biased stochastic oracle where σ represents the stochastic noise and δ the deterministic bias. Indeed, as we have seen in Chapter 4, the notion of (δ, L) oracle allows us to consider different natural notions of bias:

- ◇ $g_{\delta,L}(x)$ **is an approximate gradient of** f
  If $f \in F_{\overline{L}}^{1,1}(Q)$, $\|\nabla f(x) - g_{\delta,L}(x)\|_* \leq \Delta$ and Q is bounded with diameter $D = \max_{x \in Q, y \in Q} \|x - y\|$ then $(f_{\delta,L}(x) = f(x) - \Delta D, g_{\delta,L}(x))$ is a (δ, L) oracle with $\delta = 2\Delta D$ and $L = \overline{L}$.

- ◇ $g_{\delta,L}(x)$ **is a gradient of** f **computed at a shifted point** $\overline{x}$
  If $f \in F_{\overline{L}}^{1,1}(Q)$ and $g_{\delta,L}(x) = \nabla f(\overline{x})$ then
  $(f_{\delta,L}(x) = f(\overline{x}) + \langle \nabla f(\overline{x}), x - \overline{x} \rangle, g_{\delta,L}(x) = \nabla f(\overline{x}))$ is a (δ, L) oracle with $\delta = \overline{L} \|x - \overline{x}\|^2$ and $L = 2\overline{L}$.

- ◇ f **is in fact nonsmooth and** $g_{\delta,L}(x)$ **is a subgradient of** f
  If f is nonsmooth with bounded variations of subgradients i.e.

$$\|g(x) - g(y)\|_* \leq M, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x), g(y) \in \partial f(y)$$

then $(f_{\delta,L}(x) = f(x), g_{\delta,L}(x) = g(x))$ is a (δ, L) oracle with δ an arbitrary positive constant and $L = \frac{M^2}{2\delta}$. (In this case, the bias δ correspond to the fact that the function is not as smooth as expected).

Remark 7.2. We use the denomination smooth convex problem for (7.1) even if the functions f and h can both be nonsmooth. The reason is the fact that component h does not play any role in the design and the convergence rate of the first-order methods that we will consider. Furthermore, function f is typically a smooth convex function with Lipschitz-continuous gradient. A nonsmooth f can be also considered but the nonsmoothness is seen in this case as a bias with respect to the desired situation (using the notion of (δ, L) oracle). This generality is not the main goal of this chapter, we are mainly interested in the minimization of a smooth convex function f endowed with stochastic oracle (augmented eventually by an easy nonsmooth convex function h).

Remark 7.3. The first-order methods developed in this chapter will use only stochastic estimates of the gradient $G_{\delta,L}(x_i, \xi_i)$ at different search points x_i, and not the corresponding estimates of the function value. We need $F_{\delta,L}(x, \xi)$, only when we want to estimate the quality of a point $x \in Q$ for the objective function (see section 7.7).

7.1.2 Examples

Before developing different stochastic first-order methods, we present some examples of problems of the form (7.1) equipped with a stochastic oracle.

Lasso problem with stochastic gradient

The Lasso problem corresponds to problem (7.1) with $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $h(x) = \lambda \|x\|_1$ with $\lambda > 0$ and $Q = \mathbb{R}^n$. When using the Euclidean setup, the sparsity promoter $h(x) = \lambda \|x\|_1$ can be considered as an easy convex function. Indeed for all $g \in \mathbb{R}^n$ and $\lambda, \beta \in \mathbb{R}_0^+$, we have

$$\arg\min_{x \in \mathbb{R}^n} \{\langle g, x\rangle + \lambda \|x\|_1 + \frac{\beta}{2}\|x - z\|_2^2\} = \tau_{\frac{\lambda}{\beta}}(z - \frac{1}{\beta}g)$$

where $\tau_\alpha(x)^i = (|x^i| - \alpha)_+ \text{sgn}(x^i)$ is the shrinkage operator.
We are interested in situations where $\nabla f(x)$ is not computed exactly.

- ◇ One possible situation is when the computation of $\nabla f(x)$ is really affected by a stochastic noise and a bias. This is the case for example when instead of computing $\nabla f(x) = A^T Ax - A^T b$, we are only able to compute $G_{\delta,L}(x, \xi) = A^T A\overline{x} - A^T b + \xi$ where

1. ξ is a stochastic perturbation such that $E[\xi] = 0$ and $E[\|\xi\|_2^2] \leq \sigma^2$
2. $\overline{x}$ is shifted w.r.t. x such that $\|x - \overline{x}\|_2^2 \leq \frac{\delta}{\lambda_{\max}(A^T A)}$ (see subsection 4.1.3 in Chapter 4).

$\diamond$ Another situation is when the stochasticity is not present in the problem initially but we introduce it in order to reduce the computational cost of the first-order information. In the Lasso problem, introducing a randomization can be interesting for example when the number of rows N of A is very large. In this case, denoting by a_i the ith row of A, the computation of the exact gradient $\nabla f(x) = \sum_{i=1}^{N}(x^T a_i - b_i)a_i$ can be very expensive ($O(nN)$ basics operations). It can be interesting to replace $\nabla f(x)$ by an unbiased estimate $G_{0,L}(x,\xi) = \frac{N}{M}\sum_{j=1}^{M}(x^T a_{\xi_j} - b_{\xi_j})a_{\xi_j}$ where $\{\xi_1, ..., \xi_M\}$ is a subset of rows uniformly chosen from $\{1, ..., N\}$. When M is chosen significantly smaller than N, the computation of this stochastic gradient is of course cheaper. However, replacing the exact gradient by this stochastic estimate introduces a stochastic noise σ that depends on dissimilarities between different rows of A.

Smooth Expectation function

Let X be a random vector supported on $\Xi \subset \mathbb{R}^d$. Assume that f itself is defined by an expectation:

$$f(x) = E_{\eta\sim X}[F(x,\eta)] = \int_{\Xi} F(x,\eta)dP(\eta),$$

where $F(.,\eta) \in F^{1,1}_{L(\eta)}(Q)$ for almost all $\eta \in \Xi \subset \mathbb{R}^d$. Then we have $\nabla f(x) = E_{\eta\sim X}[\nabla_1 F(x,\eta)]$ (see [73]) and $f \in F^{1,1}_{L}(Q)$ where $L = \int_{\Xi} L(\eta)dP(\eta)$ (assuming that $L(.)$ is integrable on Ξ i.e. that $L < \infty$). However the computation of $\nabla f(x)$ i.e. of a multidimensional integral is too costly when the dimension d is high. Therefore it is typical to replace $\nabla f(x)$ by a stochastic gradient: we sample from the distribution of X, obtaining $\xi \in \Xi$ and compute $G_{0,L}(x,\xi) = \nabla_1 F(x,\xi)$. This stochastic gradient is unbiased (i.e. $\delta = 0$): $E_{\xi\sim X}[G_{\delta,L}(x,\xi)] = \nabla f(x)$ and the noise that we introduce can be characterized by

$$\begin{aligned}\sigma^2 &= E_{\xi\sim X}[(\|\nabla f(x) - G_{0,L}(x,\xi)\|_E^*)^2] \\ &= \int_{\Xi}\left\|\left(\int_{\Xi}(\nabla_1 F(x,\eta) - \nabla_1 F(x,\xi))dP(\eta)\right\|_E^*\right)^2 dP(\xi).\end{aligned}$$

Of course, we can also add to f an easy convex function h, such as a sparsity promoter $h(x) = \lambda\|x\|_1$.

Randomization of Quadratic Problem

We consider the situation where

1. $f(x) = l(x) + x^T Ax$ with $l \in F_{L_l}^{1,1}(Q)$ and $A \succeq 0$
2. $h(x) = 0$
3. $Q = \Delta_n = \{x \in \mathbb{R}_+^n : \sum_{i=1}^n x^i = 1\}$.

For such a problem on the simplex, it is natural to use the l_1 setup (it does not mean that we add to f the l_1 norm, $h(x) = \lambda \|x\|_1$, but only that we use $\|.\|_E = \|.\|_1$ and the entropy prox-function).

When the problem size is very large and when the computational cost of ∇h is not too expensive, the matrix vector product Ax becomes the dominant cost in the computation of $\nabla f(x)$. It could be very interesting to replace the costly matrix vector product by a randomized one. One possibility is to pick up from A the column i with probability x^i and to consider Ae_i (i.e. the $i$the column of A) as the stochastic estimate of Ax. This randomization technique for matrix-vector multiplication on the unit simplex has been introduced recently in [34]. The obtained oracle is unbiased (i.e. $\delta = 0$) and introduces a noise of order $\|A\|_\infty$ that can be reasonable when $L_l >> \|A\|_\infty$.

7.2 Stochastic Primal Gradient Method

7.2.1 Scheme

In this method, we use only one sequence of coefficients $\{\beta_k\}_{k\geq 0}$. We assume that $\beta_k > L$ for all $k \geq 0$ and denote $\gamma_k = \frac{1}{\beta_k}$ (that can be interpreted as the stepsize).

Algorithm 23 Stochastic Primal Gradient Method (SPGM)

1: Compute $x_0 = \arg\min_{x\in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Let ξ_k be a realization of the random variable X_k
4: Compute $G_{\delta,L}(x_k, \xi_k)$
5: Compute $x_{k+1} = \arg\min_{x\in Q}[\langle G_{\delta,L}(x_k, \xi_k), x - x_k\rangle + h(x) + \beta_k V(x, x_k)]$
6: **end for**

When we stop the scheme, the approximate solution is constructed as a weighted

average of the search points

$$y_k = \frac{1}{\sum_{i=0}^{k-1} \gamma_i} \sum_{i=0}^{k-1} \gamma_i x_{i+1}.$$

Remark 7.4. In the literature, the stochasticity is typically assumed to enter the scheme via i.i.d. random variables. Here, we consider a more general situation where a random variable X is associated with each $x \in Q$. It means that:

1. The distribution of X_i depends only on the current iterate x_i, not on the history of the process $\xi_{[i]} = (\xi_0, ..., \xi_{i-1})$ that has led the scheme to the point x_i
2. The random variables $X_0, ..., X_k$ can have different distributions but must satisfy the uniform bounds 7.3, 7.4 and 7.5 with the same σ and the same δ.

Of course, if we consider the particular case where all the random variables X have the same distribution, independently of x, we come back to the i.i.d. case. We will only use this i.i.d. assumption in the Section 7.6 and 7.7 in order to develop probabilities of large deviations.

The Primal Gradient Method is the most natural, classical first-order method. In the deterministic smooth case, when the Euclidean setup is used and $h = 0$ we retrieve the classical gradient method (see [58]):

$$\begin{aligned} x_{k+1} &= \arg\min_{x \in Q} \{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\beta_k}{2} \|x - x_k\|_2^2\} \\ &= \pi_Q(x_k - \frac{1}{\beta_k} \nabla f(x_k)) \end{aligned}$$

where π_Q denotes the Euclidean projection on Q.

If we choose all the coefficients β_i equal to the the Lipschitz constant of the gradient L, we obtain the famous convergence rate $O\left(\frac{LR^2}{k}\right)$ (which is however non-optimal for smooth convex problems).

This family of schemes has also attracted a lot of attention in nonsmooth convex optimization, as it is simply the subgradient method ([74]) if we use the Euclidean setup and the Mirror Descent method ([55, 3]) with a general setup. With an increasing sequence of coefficients $\beta_i = \Theta\left(\frac{M\sqrt{i}}{R}\right)$, we obtain the optimal convergence rate $O\left(\frac{MR}{\sqrt{k}}\right)$ for deterministic nonsmooth convex probem where M denotes the Lipschitz-constant of the function.

In stochastic nonsmooth convex optimization, this scheme corresponds to the Stochastic Approximation (SA) method in the Euclidean case and to the Mirror Descent Stochastic Approximation (MDSA) method ([52]) in the general case. With the same kind of decreasing stepsizes γ_i (i.e. of increasing coefficients β_i) than in the deterministic case, these methods reach the unimprovable convergence rate $O\left(\frac{MR}{\sqrt{k}} + \frac{\sigma R}{\sqrt{k}}\right)$ where σ denotes the stochastic noise of the oracle.

In stochastic smooth convex optimization, this scheme has been considered recently by Lan in [44] under the name of Modified Mirror Descent SA method (MMDSA). He proposes to construct the approximate solution (i.e. the point for which we have the convergence rate) as $\frac{1}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\gamma_i x_{i+1}$ (instead of $\frac{1}{\sum_{i=0}^{k}\gamma_i}\sum_{i=0}^{k}\gamma_i x_i$ for the usual MDSA method) and a constant stepsize policy based on the oracle noise σ and on the performed number of iterations k. This method exhibits the rate of convergence $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ which is optimal with respect to the stochastic noise σ but not with respect to L, the Lipschitz-constant of the gradient.

In this section, we generalize the result of Lan in three directions:

- ◇ We consider the biased case, when the expectation of the stochastic gradient $G_{\delta,L}(x,\xi)$ is itself affected by a deterministic error δ. In this case, the convergence rate is $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$
- ◇ We propose a new stepsize policy that does not need anymore the knowledge of the performed number of iterations and gives the same convergence rate (up to a logarithmic factor)
- ◇ We consider the composite case when we add to f an easy convex function h (possibly nonsmooth) i.e. that can be kept without modification in the auxiliary subproblems.

Let us start with the general convergence rate of this Stochastic Primal Gradient Method (SPGM):

7.2.2 General Convergence Rate

Theorem 7.1. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then, if $\beta_i > L$ for all $i \geq 0$, the sequence y_k generated by the Stochastic Primal Gradient Method, when applied to the composite*

function ϕ, satisfies

$$\begin{aligned}\phi(y_k) - \phi^* &\leq \frac{1}{\sum_{i=0}^{k-1}\gamma_i}(V(x^*, x_0) + \sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}(\|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|_E^*)^2 \\ &+ \sum_{i=0}^{k-1}\gamma_i\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i\rangle) + \delta.\end{aligned}$$

Proof. For simplicity, in all proofs of this chapter, we denote $f_i = f_{\delta,L}(x_i)$, $F_i = F_{\delta,L}(x_i, \xi_i)$, $g_i = g_{\delta,L}(x_i)$ and $G_i = G_{\delta,L}(x_i, \xi_i)$.
Let $gh(x_{k+1}) \in \partial h(x_{k+1})$, from the definition of x_{k+1}, we have:

$$\langle \gamma_k G_k + \gamma_k gh(x_{k+1}) + \nabla d(x_{k+1}) - \nabla d(x_k), u - x_{k+1}\rangle \geq 0, \quad \forall u \in Q.$$

When rearranging terms, this inequality can be written as:

$$\begin{aligned}\gamma_k\langle G_k, x_k - u\rangle &\leq V(u, x_k) - V(u, x_{k+1} + \gamma_k\langle G_k, x_k - x_{k+1}\rangle \\ &\quad - V(x_{k+1}, x_k) + \gamma_k\langle gh(x_{k+1}), u - x_{k+1}\rangle.\end{aligned}$$

Denoting $d_k = \gamma_k\langle G_k, x_k - x_{k+1}\rangle - V(x_{k+1}, x_k)$, we obtain:

$$\begin{aligned}d_k &\overset{(2.10)}{\leq} \gamma_k\langle G_k, x_k - x_{k+1}\rangle - \frac{1}{2}\|x_k - x_{k+1}\|_E^2 \\ &= \gamma_k[\langle g_k, x_k - x_{k+1}\rangle - \frac{L}{2}\|x_k - x_{k+1}\|_E^2] \\ &\quad + \gamma_k[\langle G_k - g_k, x_k - x_{k+1}\rangle - \frac{\beta_k - L}{2}\|x_k - x_{k+1}\|_E^2] \\ &\overset{(7.2)}{\leq} \gamma_k[f_k - f(x_{k+1}) + \delta] + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2.\end{aligned}$$

where we use in the last inequality the fact that for all $g \in E^*, x \in E, \gamma > 0$:

$$\langle g, x\rangle - \frac{\zeta}{2}\|x\|_E^2 \leq \frac{1}{\zeta}(\|g\|_E^*)^2. \tag{7.6}$$

Therefore, we obtain:

$$\begin{aligned}\gamma_k\langle G_k, x_k - u\rangle &\leq V(u, x_k) - V(u, x_{k+1}) + \gamma_k[f_k - f(x_{k+1}) + \delta] \\ &\quad + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2 + \gamma_k\langle gh(x_{k+1}), u - x_{k+1}\rangle \\ &\leq V(u, x_k) - V(u, x_{k+1}) + \gamma_k[f_k - f(x_{k+1}) + \delta] \\ &\quad + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2 + \gamma_k(h(u) - h(x_{k+1})).\end{aligned}$$

i.e:

$$\begin{aligned} & \gamma_k[f(x_{k+1}) + h(x_{k+1})] \\ \leq \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k[f_k + \langle g_k, u - x_k\rangle] \\ & + \gamma_k\langle G_k - g_k, u - x_k\rangle + \gamma_k\delta + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2 + \gamma_k h(u) \\ \overset{(7.2)}{\leq} \quad & V(u, x_k) - V(u, x_{k+1}) + \gamma_k\phi(u) + \gamma_k[\langle G_k - g_k, u - x_k\rangle] \\ & + \gamma_k\delta + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2. \end{aligned}$$

In particular, choosing $u = x^*$:

$$\begin{aligned} \gamma_k\phi(x_{k+1}) \quad \leq \quad & V(x^*, x_k) - V(x^*, x_{k+1}) + \gamma_k\phi^* + \gamma_k[\langle G_k - g_k, x^* - x_k\rangle] \\ & + \gamma_k\delta + \frac{\gamma_k}{\beta_k - L}(\|G_k - g_k\|_E^*)^2. \end{aligned}$$

Summing these inequalities, we obtain:

$$\begin{aligned} \sum_{i=0}^{k-1}\gamma_i(\phi(x_{i+1}) - \phi^*) \quad \leq \quad & V(x^*, x_0) + \sum_{i=0}^{k-1}\gamma_i\langle G_i - g_i, x^* - x_i\rangle \\ & + \sum_{i=0}^{k-1}\gamma_i\delta + \sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}(\|G_i - g_i\|_E^*)^2 \end{aligned}$$

and therefore:

$$\begin{aligned} \phi(y_k) - \phi^* \quad \leq \quad & \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\left(V(x^*, x_0) + \sum_{i=0}^{k-1}\gamma_i\langle G_i - g_i, x^* - x_i\rangle\right) \\ & + \frac{1}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}(\|G_i - g_i\|_E^*)^2 + \delta. \end{aligned}$$

□

Remark 7.5. We observe that the convergence rate does not depend on the difference $(F_{\delta,L}(x_i, \xi_i) - f_{\delta,L}(x_i))$. This is natural since the scheme itself does not use $F_{\delta,L}(x_i, \xi_i)$, the stochastic estimate of the function value. This property is shared by all methods considered in this chapter.

Taking now the expectation with respect to the history of the random process $\xi_{[i]} = (\xi_0, ..., \xi_i)$, we obtain the following result:

Theorem 7.2. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then, if $\beta_i > L$ for all $i \geq 0$, the Stochastic*

Primal Gradient Methods, when applied to the composite function ϕ, exhibits on average the convergence rate

$$E_{\xi_1 \sim X_1, \ldots \xi_k \sim X_k}[\phi(y_k) - \phi^*] \leq \frac{V(x^*, x_0)}{\sum_{i=0}^{k-1} \gamma_i} + \frac{1}{\sum_{i=0}^{k-1} \gamma_i} \sum_{i=0}^{k-1} \frac{\gamma_i}{\beta_i - L} \sigma^2 + \delta.$$

Proof. As $E_{\xi_i \sim X_i}[G_i | \xi_{[i-1]}] = g_i$ and as x_i is a deterministic function of $(\xi_1, ..., \xi_{i-1})$, the expectation of $\langle G_i - g_i, x^* - x_i \rangle$, conditional on $\xi_{[i-1]} = (\xi_1, ..., \xi_{i-1})$, is zero. Therefore, we have $E_{\xi_1 \sim X_1, \ldots, \xi_k \sim X_k}[\sum_{i=0}^{k-1} \gamma_i \langle G_i - g_i, x^* - x_i \rangle] = 0$. Furthermore, by assumption, $E_{\xi_i \sim X_i}[\|G_i - g_i\|_*^2 | \xi_{[i-1]}] \leq \sigma^2$ and we obtain: $E_{\xi_1 \sim X_1, \ldots, \xi_k \sim X_k}[\sum_{i=0}^{k-1} \frac{\gamma_i}{\beta_i - L} \|G_i - g_i\|_*^2] \leq \sum_{i=0}^{k-1} \frac{\gamma_i}{\beta_i - L} \sigma^2$. □

7.2.3 Choice of Stepsizes

Why do we need new stepsizes rules ?

In the deterministic smooth case (i.e. when the function f is smooth with a Lipschitz continuous gradient and the oracle is exact), the optimal stepsize (see [58]) is constant and equal to the inverse of the Lipschitz-constant of the gradient: $\gamma_i = \frac{1}{L}, \quad \forall i \geq 0$. If we keep this stepsizes rule in the stochastic case, we cannot apply Theorem 7.1 (that assumes $\gamma_i < \frac{1}{L}$) but with an easy modification in the proof of this theorem, we can obtain the following upper-bound:

$$\phi(y_k) - \phi^* \leq \frac{1}{\sum_{i=0}^{k-1} \gamma_i} \left(V(x^*, x_0) + \sum_{i=0}^{k-1} \gamma_i [\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_{i+1} \rangle] \right) + \delta.$$

But as x_{i+1} depends on $G_{\delta,L}(x_i, \xi_i)$, we cannot say that

$$E[\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_{i+1} \rangle | \xi_{[i-1]}] = 0$$

but only that:

$$\begin{aligned} & E[\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_{i+1} \rangle | \xi_{[i-1]}] \\ \leq \; & \sqrt{E[\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_*^2 | \xi_{[i-1]}]} D \\ \leq \; & \sigma D \end{aligned}$$

where $D = \max_{x \in Q, y \in Q} \|x - y\|$ is the diameter of the feasible set. Therefore we obtain

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{2k} + \delta + \sigma D$$

where R is an upper bound on $\sqrt{2V(x^*, x_0)}$.
We see that with the classical stepsize policy, the effect of the stochastic noise

does not decrease with the iterations. This is a behavior that we want to avoid, since it would be preferable to obtain a method that could converge to the optimal value of our problem ϕ^* (or at least to $\phi^* + \delta$ in the biased case).

If we consider $\gamma_i = \frac{1}{CL}$ with $C > 1$, in this case we can apply the Theorems 7.1 and 7.2 and obtain:

$$E[\phi(y_k) - \phi^*] \leq \frac{CLR^2}{2k} + \delta + \frac{\sigma^2}{(C-1)L}$$

but here also we obtain the same kind of behavior with a method that cannot decrease the stochastic noise effect when we increase the number of iterations. If we want to be able to converge to ϕ^* in the unbiased case or to $\phi^* + \delta$ in the biased case, a decreasing stepsize policy must be used.

Remark 7.6. For nonsmooth problems, the same kind of decreasing stepsize $\gamma_i = O\left(\frac{R}{M\sqrt{i}}\right)$ can be used both in the deterministic and the stochastic case. For smooth problems, the more aggressive constant stepsize $\gamma_i = O\left(\frac{1}{L}\right)$ (that leads to the improvement of the convergence rate in the deterministic case from $O\left(\frac{1}{\sqrt{k}}\right)$ to $O\left(\frac{1}{k}\right)$) is too large and not able to decrease the stochastic noise. In some sense, the gradient method is faster than the subgradient method but more sensible with respect to the stochastic error σ. When stochasticity is present, we need to consider decreasing stepsize also in the smooth case (but decreasing only in term of σ not of L, i.e of the form $O\left(\frac{1}{L+\frac{\sigma}{R}\sqrt{i}}\right)$).

A new stepsize rule

By the complexity theory of first-order methods (see [55, 52, 44]), the best that we can expect in the stochastic case is a method that reduces the noise effect σ by a quantity $\Theta(\frac{\sigma R}{\sqrt{k}})$ after k iterations. This result gives us possibility to expect a better behavior for the SPGM that what we have obtained using the classical constant stepsize in the last section. In the same time, there is no hope to obtain a method with convergence rate $\Theta(\frac{LR^2+\sigma R}{k} + \delta)$. If we assume that the number of iterations N is known in advance, we can obtain the rate $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ relatively easily. A constant stepsize (but that depends on the performed number of iterations) can be chosen. In [44], Lan has proposed the rule $\gamma_i = \min\left(\frac{1}{2L}, \sqrt{\frac{R^2}{2N\sigma^2}}\right), \quad \forall i \geq 0$ and obtained this desired rate of convergence. Another possible choice is $\gamma_i = \frac{1}{L+\frac{\sigma}{R}\sqrt{N}}$ that leads to

$$E[\phi(y_N) - \phi^*] \leq \frac{LR^2}{2N} + \frac{3\sigma R}{2\sqrt{N}} + \delta.$$

Remark 7.7. For a first-order method with convergence rate $O\left(\frac{LR^2}{k}\right)$ in the deterministic exact case, the effect of the deterministic bias δ cannot be better than an additional term δ (see Theorem 4.12). Therefore, this convergence rate has an optimal dependence in δ and σ.

Remark 7.8. It is possible to obtain a better dependence in L using an accelerated method, like the Stochastic Fast Gradient Method (SFGM) (see Section 7.5) with convergence rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}+k\delta\right)$ but, in this case, we pay this acceleration by a unavoidable worst dependence in δ.

However, the need of fixing in advance the number of iterations is not really a desirable property. Often in practice, we want to run a method for a given time and not for a given number of iterations. For this reason, it is interesting to develop a practical stepsizes rule which is not based on an a priori knowledge of the performed number of iterations and at the same times that keeps the convergence rate $\Theta(\frac{LR^2}{k}+\frac{\sigma R}{\sqrt{k}}+\delta)$. This is not trivial. Indeed, contrarily to more sophisticated methods of the latter sections, there is not a lot of freedom in the SPGM: we have only one sequence of coefficients β_i $(=\frac{1}{\gamma_i})$. Consider the choice:

$$\gamma_i = \frac{L+\frac{\sigma}{2R}\sqrt{i+1}}{(L+\frac{\sigma}{R}\sqrt{i+1})^2}.$$

This stepsize decreases with rate $\Theta\left(\frac{1}{L+\frac{\sigma}{R}\sqrt{i}}\right)$ and we retrieve the optimal stepsize $\gamma_i=\frac{1}{L}$ in the deterministic case. We have for all $k\geq 1$:

$$\begin{aligned}
\sum_{i=0}^{k-1}\gamma_i &= \sum_{i=1}^{k}\frac{L+\frac{\sigma}{2R}\sqrt{i}}{(L+\frac{\sigma}{R}\sqrt{i})^2}\\
&\geq \int_1^{k+1}\frac{L+\frac{\sigma}{2R}\sqrt{x}}{(L+\frac{\sigma}{R}\sqrt{x})^2}dx = \left[\frac{x}{L+\frac{\sigma}{R}\sqrt{x}}\right]_1^{k+1} = \frac{Lk+\frac{\sigma}{R}(k+1-\sqrt{k+1})}{(L+\frac{\sigma}{R})(L+\frac{\sigma}{R}\sqrt{k+1})}\\
&\geq \frac{(2-\sqrt{2})\frac{\sigma}{R}k+Lk}{(L+\frac{\sigma}{R})(L+\frac{\sigma}{R}\sqrt{k+1})} \geq \frac{(2-\sqrt{2})k}{L+\frac{\sigma}{R}\sqrt{k+1}}
\end{aligned}$$

and therefore $\frac{1}{\sum_{i=0}^{k-1}\gamma_i}\leq\frac{L+\frac{\sigma}{R}\sqrt{k+1}}{(2-\sqrt{2})k}$. On the other hand, we have:

$$\begin{aligned}
\frac{\gamma_i}{\beta_i-L} &= \frac{(L+\frac{\sigma}{2R}\sqrt{i+1})^2}{(L+\frac{\sigma}{R}\sqrt{i+1})^2(\frac{\sigma^2}{R^2}(i+1)+\frac{3}{2}\frac{L\sigma}{R}\sqrt{i+1})}\\
&\leq \frac{1}{\frac{\sigma^2}{R^2}(i+1)+\frac{3}{2}\frac{L\sigma}{R}\sqrt{i+1}} \leq \frac{1}{\frac{\sigma^2}{R^2}(i+1)}.
\end{aligned}$$

and therefore $\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i-L}\leq\frac{R^2}{\sigma^2}Har(k)$ where $Har(k)=\sum_{i=1}^{k}\frac{1}{i}\leq 1+\ln(k)$.

We obtain finally the convergence rate:

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{(2-\sqrt{2})k}(Har(k) + \frac{1}{2}) + \frac{\sigma\sqrt{k+1}R}{(2-\sqrt{2})k}(Har(k) + \frac{1}{2}) + \delta.$$

As $Har(k) \leq 1 + \ln(k)$, we retrieve, up to a logarithmic factor, a rate of the form $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ but now using varying stepsizes that do not assume the knowledge of the performed number of iterations.

Remark 7.9. If we want to avoid the logarithmic factor in the convergence rate and if the set Q is bounded with diameter D, we can define the approximate solution as: $\overline{y}_N : \frac{1}{\sum_{i=N/2-1}^{N-1}\gamma_i}\sum_{i=N/2-1}^{N-1}\gamma_i x_{i+1}$ averaging only the last $\frac{N}{2}$ search points x_i (for simplicity, we assume here that N is even). In this case we obtain: $E[\phi(\overline{y}_N) - \phi^*] \leq \frac{2\sqrt{2}}{2-\sqrt{2}}\left(1 + \frac{2}{N} + \ln(2)\right)\left(\frac{LD^2}{N} + \frac{\sigma D}{\sqrt{N}}\right)$. However this choice of averaging assumes the storage of all test points in memory, when N is not known a priori.

7.3 Stochastic estimate functions

In this chapter, we generalize the concept of estimate functions sequence, presented in subsection 2.4.3 assuming now that the model of the function $\Psi_k(x)$ is constructed using stochastic first-order information (possibly with bias) and the sequences $\{y_k\}_{k\geq 0}$ and $\{\Psi_k(x)\}_{k\geq 0}$ satisfies the two inequalities:

$$A_k\phi(y_k) \leq \Psi_k^* + E_k \text{ and } \Psi_k(x) \leq A_k\phi(x) + \beta_k d(x) + \overline{E}_k(x), \quad \forall x \in Q$$

where E_k and $\overline{E}_k(x)$ represent random errors coming from the stochastic noise σ and the bias δ.
With this notion of stochastic estimate functions, we obtain the convergence rate:

$$\phi(y_k) - \phi^* \leq \frac{\beta_k d(x^*)}{A_k} + \frac{E_k + \overline{E}_k(x^*)}{A_k}$$

and therefore:

$$E[\phi(y_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{E[E_k + \overline{E}_k(x^*)]}{A_k}$$

since the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ are deterministic (they will be based on the noise level σ but not on the realizations of the random variables $X_1, ..., X_k$).

Using this framework, we will develop in Section 7.4 a stochastic dual gradient method with convergence rate $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ and in Section 7.5, a stochastic fast gradient method with convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta\right)$.

Remark 7.10. In the deterministic case, the model $\Psi_k(x)$ is typically chosen of the form: $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^{k} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + h(x)]$. In the stochastic case, we will simply modify this model using the stochastic first-order information instead of the exact one: $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^{k} \alpha_i [F_{\delta,L}(x_i, \xi_i) + \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + h(x)]$.

Remark 7.11. Compared with the primal gradient method, the methods based on this principle of estimate functions are more sophisticated and often less intuitive.

However, they provide us typically with more degrees of freedom (multiple sequences of coefficients, multiple sequences of iterates), that make these methods more flexible for new situations (we will see that adaptation of the DGM and FGM to the stochastic case is in some sense easier than for the PGM) and well-suited for acceleration (cf. the optimal rate of the Fast Gradient method.)

7.4 Stochastic Dual Gradient Method

7.4.1 Scheme

In this method we use two sequences of coefficients:

$$\{\alpha_k\}_{k \geq 0} \text{ with } \alpha_0 \in]0, 1] \text{ and } \{\beta_k\}_{k \geq 0} \text{ with } \beta_{k+1} \geq \beta_k > L \quad \forall k \geq 0.$$

Furthermore the two sequences must satisfy the coupling condition:

$$\beta_k \geq \alpha_{k+1} \beta_{k+1}, \quad \forall k \geq 0. \tag{7.7}$$

We define also $A_k = \sum_{i=0}^{k} \alpha_i$.

Algorithm 24 Stochastic Dual Gradient Method (SDGM)

1: Compute $x_0 = \arg\min_{x \in Q} d(x)$
2: Let ξ_0 be a realization of the random variable X_0
3: Compute $G_{\delta,L}(x_0, \xi_0)$
4: Compute

$$w_0 = \arg\min_{x \in Q} \{\beta_0 d(x) + \alpha_0 \langle G_{\delta,L}(x_0, \xi_0), x - x_0 \rangle + \alpha_0 h(x)\} \tag{7.8}$$

5: **for** k = 0 : ... **do**
6: Compute

$$x_{k+1} = \arg\min_{x \in Q} \{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + A_k h(x)\} \tag{7.9}$$

7: Let ξ_{k+1} be a realization of the random variable X_{k+1}
8: Compute $G_{\delta,L}(x_{k+1}, \xi_{k+1})$
9: Compute

$$w_{k+1} = \arg\min_{x \in Q} \{\beta_{k+1} V(x, x_{k+1}) + \langle G_{\delta,L}(x_{k+1}), x - x_{k+1} \rangle + h(x)\} \tag{7.10}$$

10: **end for**

When we stop the scheme, the approximated solution is calculated as a weighted sum of the points $\{w_i\}_{0 \leq i \leq k}$:

$$y_k = \frac{1}{\sum_{i=0}^{k} \alpha_i} \sum_{i=0}^{k} \alpha_i w_i.$$

The dual gradient method has been introduced in [62] by Nesterov in the deterministic (composite) case and using the Euclidean setup.
We generalize this method in two directions:

- ◇ We generalize the method to the non-Euclidean setting, using auxiliary subproblems based only on the prox-function $d(x)$
- ◇ We adapt the method to the stochastic case (possibly with bias). We will see that the classical choice $\beta_i = L$ is not anymore a good idea when stochasticity is present and we propose an increasing policy for the sequence $\{\beta_i\}$ that leads to a convergence rate of the form $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ (or $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right)$ when the oracle is biased).

But first, we start with the general convergence rate of this stochastic dual gradient method:

7.4.2 General Convergence rate

Denote by $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^{k} \alpha_i [F_{\delta,L}(x_i, \xi_i) + \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + h(x)]$, our model of the objective function, $\Psi_k^* = \min_{x \in Q} \Psi_k(x)$ its minimal value on the feasible set and $\xi_{[k]} = (\xi_0, ..., \xi_k)$ the history of the random process after k iterations.

Let us show that the two sequences $\{y_k\}_{k \geq 0}$ and $\{\Psi_k(x)\}_{k \geq 0}$ define a sequence of estimate functions.

Lemma 7.3. *For all $k \geq 0$, we have:*

1.

$$A_k \phi(y_k) \leq \Psi_k^* + E_k \tag{7.11}$$

where

$$\begin{aligned} E_k &= \sum_{i=0}^{k} \alpha_i \delta + \sum_{i=0}^{k} \alpha_i [f_{\delta,L}(x_i) - F_{\delta,L}(x_i, \xi_i)] \\ &\quad + \sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L} (\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_E^*)^2 \end{aligned}$$

2.

$$\Psi_k(x) \leq A_k \phi(x) + \beta_k d(x) + \overline{E}_k(x), \quad \forall x \in Q \tag{7.12}$$

where $\overline{E}_k(x) = \sum_{i=0}^{k} \alpha_i [F_{\delta,L}(x_i, \xi_i) - f_{\delta,L}(x_i) + \langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x - x_i \rangle]$

Proof. 1. First, we will show by recurrence that the inequality: $\sum_{i=0}^{k} \alpha_i \phi(w_i) \leq \Psi_k^* + E_k$ is satisfied for all $k \geq 0$.

- For $k = 0$, we have:

$$\begin{aligned} \phi(w_0) &\overset{(7.2)}{\leq} f_0 + \langle g_0, w_0 - x_0 \rangle + \frac{L}{2} \|w_0 - x_0\|_E^2 + \delta + h(w_0) \\ &= F_0 + \langle G_0, w_0 - x_0 \rangle + \frac{\beta_0}{2} \|w_0 - x_0\|_E^2 + \delta + h(w_0) \\ &\quad + (f_0 - F_0) + \langle g_0 - G_0, w_0 - x_0 \rangle - \frac{\beta_0 - L}{2} \|w_0 - x_0\|_E^2 \end{aligned}$$

and we obtain, using (2.8), (7.6) and the fact $0 < \alpha_0 \leq 1$:

$$\begin{aligned}\alpha_0\phi(w_0) &\leq \alpha_0[F_0 + \langle G_0, w_0 - x_0\rangle + h(w_0)] + \beta_0 d(w_0)\\ &\quad + \frac{\alpha_0}{\beta_0 - L}(\|G_0 - g_0\|_E^*)^2 + \alpha_0(f_0 - F_0)\\ &\overset{(7.8)}{=} \Psi_0^* + \frac{\alpha_0}{\beta_0 - L}(\|G_0 - g_0\|_E^*)^2 + \alpha_0(f_0 - F_0).\end{aligned}$$

• Now assume that this inequality is satisfied for $k \geq 0$ i.e. that we have

$$\sum_{i=0}^{k} \alpha_i \phi(w_i) \leq \Psi_k^* + E_k.$$

Then as $\beta_{k+1} \geq \beta_k$ and by definition of V we have:

$$\begin{aligned}\Psi_{k+1}^* &= \min_{x\in Q}\{\beta_{k+1}d(x) + \sum_{i=0}^{k+1}\alpha_i[F_i + \langle G_i, x - x_i\rangle + h(x)]\}\\ &\geq \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \beta_k d(x_{k+1}) + \beta_k\langle\nabla d(x_{k+1}), x - x_{k+1}\rangle\\ &\quad + \sum_{i=0}^{k+1}\alpha_i[F_i + \langle G_i, x - x_i\rangle]\} + A_{k+1}h(x).\end{aligned}$$

Let $gh(x_{k+1}) \in \partial h(x_{k+1})$, by optimality condition defining x_{k+1}:

$$\langle\beta_k\nabla d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i G_i + A_k gh(x_{k+1}), x - x_{k+1}\rangle \geq 0, \quad \forall x \in Q$$

and therefore:

$$\begin{aligned}\Psi_{k+1}^* &\geq \beta_k d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i[F_i + \langle G_i, x_{k+1} - x_i\rangle]\\ &\quad + \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle]\}\\ &\quad + A_{k+1}h(x) + A_k\langle gh(x_{k+1}), x_{k+1} - x\rangle\\ &\geq \beta_k d(x_{k+1}) + \sum_{i=0}^{k}\alpha_i[F_i + \langle G_i, x_{k+1} - x_i\rangle]\\ &\quad + \min_{x\in Q}\{\beta_k V(x, x_{k+1}) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle]\}\\ &\quad + A_k h(x_{k+1}) + \alpha_{k+1}h(x)\\ &= \Psi_k^* + \alpha_{k+1}\min_{x\in Q}\{F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle\\ &\quad + h(x) + \frac{\beta_k}{\alpha_{k+1}}V(x, x_{k+1})\}.\end{aligned}$$

Using the condition $\frac{\beta_k}{\alpha_{k+1}} \geq \beta_{k+1}$, the definition of w_{k+1} and the inequality $V(w_{k+1}, x_{k+1}) \geq \frac{1}{2}\|x_{k+1} - w_{k+1}\|_E^2$, we obtain

$$\begin{aligned}
\Psi_{k+1}^* &\overset{(7.7)}{\geq} \Psi_k^* + \alpha_{k+1}\min_{x\in Q}\{F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle \\
&\quad + h(x) + \beta_{k+1}V(x, x_{k+1})\} \\
&\overset{(7.10)}{=} \Psi_k^* + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, w_{k+1} - x_{k+1}\rangle \\
&\quad + h(w_{k+1}) + \beta_{k+1}V(w_{k+1}, x_{k+1})] \\
&\overset{(2.10)}{\geq} \Psi_k^* + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, w_{k+1} - x_{k+1}\rangle \\
&\quad + h(w_{k+1}) + \frac{\beta_{k+1}}{2}\|x_{k+1} - w_{k+1}\|_E^2]
\end{aligned}$$

$$\begin{aligned}
&= \Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, w_{k+1} - x_{k+1}\rangle \\
&\quad + h(w_{k+1}) + \frac{L}{2}\|w_{k+1} - x_{k+1}\|_E^2] + \alpha_{k+1}[F_{k+1} - f_{k+1} \\
&\quad + \langle G_{k+1} - g_{k+1}, w_{k+1} - x_{k+1}\rangle + \frac{\beta_{k+1} - L}{2}\|w_{k+1} - x_{k+1}\|_E^2] \\
&\overset{(7.2),(7.6)}{\geq} \Psi_k^* + \alpha_{k+1}(f(w_{k+1}) - \delta + h(w_{k+1})) \\
&\quad + \alpha_{k+1}[F_{k+1} - f_{k+1}] - \frac{\alpha_{k+1}}{\beta_{k+1} - L}(\|G_{k+1} - g_{k+1}\|_E^*)^2 \\
&\geq \sum_{i=0}^{k+1}\alpha_i(f(w_i) + h(w_i)) - E_k - \alpha_{k+1}\delta \\
&\quad + \alpha_{k+1}[F_{k+1} - f_{k+1}] - \frac{\alpha_{k+1}}{\beta_{k+1} - L}(\|G_{k+1} - g_{k+1}\|_E^*)^2
\end{aligned}$$

and therefore: $\sum_{i=0}^{k+1}\alpha_i\phi(w_i) \leq \Psi_{k+1}^* + E_{k+1}$ where $E_{k+1} = E_k + \alpha_{k+1}\delta + \alpha_{k+1}[f_{k+1} - F_{k+1}] + \frac{\alpha_{k+1}}{\beta_{k+1}-L}(\|G_{k+1} - g_{k+1}\|_E^*)^2$.

We have proved that $\sum_{i=0}^{k}\alpha_i\phi(w_i) \leq \Psi_k^* + E_k$ and using the definition of $y_k = \frac{1}{\sum_{i=0}^{k}\alpha_i}\sum_{i=0}^{k}\alpha_i w_i$, $A_k = \sum_{i=0}^{k}\alpha_i$ and the convexity of ϕ, we obtain now: $A_k\phi(y_k) \leq \Psi_k^* + E_k$ for all $k \geq 0$.

2. On the other hand, for all $x \in Q$, we also have:

$$\begin{aligned}
\Psi_k(x) &\overset{(7.2)}{\leq} \beta_k d(x) + \sum_{i=0}^{k}\alpha_i(f(x) + h(x)) \\
&\quad + \sum_{i=0}^{k}\alpha_i[F_i - f_i + \langle G_i - g_i, x - x_i\rangle] \\
&= \beta_k d(x) + A_k\phi(x) + \overline{E}_k(x). \qquad \square
\end{aligned}$$

As we have proved that we are in the framework of estimate functions, we can now obtain directly the convergence rate for the SDGM:

Theorem 7.4. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then the sequence y_k generated by the Stochastic Dual Gradient Method, when applied to the composite function ϕ, satisfies*

$$\phi(y_k) - \phi^* \leq \frac{\beta_k d(x^*)}{A_k} + \delta$$

$$+\frac{1}{A_k}\left(\sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L} \|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|_*^2 + \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i\rangle\right).$$

Taking now the expectation with respect to the random process history $\xi_{[k]}$, we obtain the following result:

Theorem 7.5. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then the Stochastic Dual Gradient Method, when applied to the composite function ϕ, exhibits on average the convergence rate*

$$E_{\xi_0 \sim X_0, \ldots \xi_k \sim X_k}[\phi(\hat{x}_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{1}{A_k}\sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L}\sigma^2 + \delta.$$

Proof. Completely similar to the proof of theorem 7.2. □

7.4.3 Choice of the Coefficients

In the deterministic smooth case, the coefficients of the dual gradient method developed in [62] are chosen constant: $\beta_i = L$ and $\alpha_i = 1$ for all $i \geq 0$.

If we keep these coefficients in the stochastic case, we cannot apply Theorem 7.4 (that assumes $\beta_i > L$) but with an easy modification in the proof of this theorem, we can obtain the following upper-bound:

$$\phi(y_k) - \phi^* \leq \frac{LR^2}{2k} + \delta + \frac{1}{k}\sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i\rangle.$$

As w_i depends on $G_{\delta,L}(x_i)$, we cannot say that $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i\rangle | \xi_{[i-1]}] = 0$ but only

$$E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - w_i\rangle | \xi_{[i-1]}]$$
$$\leq \sqrt{E[(\|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|_E^*)^2 | \xi_{[i-1]}]}D \leq \sigma D$$

where $D = \max_{x \in Q, y \in Q} \|x - y\|$ is the diameter of the feasible set. Therefore we have:

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{2k} + \delta + D\sigma.$$

We see that with the classical choice of the coefficients, the effect of the stochastic noise does not decrease with the iterations.

If we consider $\beta_i = CL$ with $C > 1$, in this case we can apply the Theorems 7.4 and 7.5 and obtain $E[\phi(y_k) - \phi^*] \leq \frac{CLR^2}{2k} + \delta + \frac{\sigma^2}{(C-1)L}$ but here also we obtain the same kind of behavior with a method that cannot decrease the stochastic noise effect when we increase the number of iterations. If we want to be able to converge to ϕ^* in the unbiased case or to $\phi^* + \delta$ in the biased case, an increasing sequence of coefficients β_i must be used.

On the other hand, often in practice, we want to run a method for a given time and not for a given number of iterations. For this reason, it is interesting to develop a practical stepsizes rule for the stochastic dual gradient method which is not based on a a priori knowledge of the performed number of iterations and at the same times that can reach the convergence rate $\Theta(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta)$.

Consider the choice $\alpha_i = a$ with $0 < a \leq 1$ and $\beta_i = L + b\frac{\sigma}{R}(i+1)^c$.
We have

- $$\frac{\beta_k R^2}{2A_k} = \frac{LR^2}{2a(k+1)} + \frac{b\sigma R}{2a(k+1)^{1-c}}$$

- $$\begin{aligned} \frac{1}{A_k}\sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L}\sigma^2 = \frac{\sigma R}{(k+1)b}\sum_{i=1}^{k+1} i^{-c} &\leq \frac{\sigma R}{(k+1)b}\int_0^{k+1} x^{-c}dx \\ &\leq \frac{\sigma R}{b(k+1)^c}. \end{aligned}$$

We obtain therefore using Theorem 7.5 :

$$E[\phi(y_k) - \phi^*] \leq \frac{LR^2}{2a(k+1)} + \frac{b\sigma R}{2a(k+1)^{1-c}} + \frac{\sigma R}{b(k+1)^c}.$$

Optimizing the rate of convergence of the term depending on σ, we choose $c = \frac{1}{2}$ for which we obtain a convergence rate of the form $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$. For the choice of a and b, we need to ensure the condition (7.7) i.e. $(L + b\frac{\sigma}{R}(k+1)^{1/2}) \geq a(L + \frac{\sigma}{R}(k+2)^{1/2})$ for all $k \geq 0$.

A sufficient condition is $a \leq \sqrt{\frac{k+1}{k+2}}$ for all $k \geq 0$ and we obtain the condition $a \leq \frac{1}{\sqrt{2}}$. We take $\alpha_i = a = \frac{1}{\sqrt{2}}$, for all $i \geq 0$ and therefore

$$E[\phi(y_k) - \phi^*] \leq \frac{\sqrt{2}LR^2}{2(k+1)} + \frac{\sqrt{2}b\sigma R}{2\sqrt{k+1}} + \frac{\sigma R}{b\sqrt{k+1}} + \delta.$$

The optimal choice for b is $2^{1/4}$ and we obtain:

Theorem 7.6. *If the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$ are chosen for all $i \geq 0$ as $\alpha_i = \frac{1}{\sqrt{2}}$ and $\beta_i = L + \frac{2^{1/4}\sigma}{R}(i+1)^{1/2}$ then the sequence generated by the SDGM satisfies:*

$$E[\phi(y_k) - \phi^*] \leq \frac{\sqrt{2}LR^2}{2(k+1)} + \frac{2^{3/4}\sigma R}{\sqrt{k+1}} + \delta = \Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right).$$

7.5 Stochastic Fast Gradient Method

7.5.1 Scheme

In this method we use also two sequences of coefficients:

$$\{\alpha_i\}_{i\geq 0} \text{ with } \alpha_0 \in]0,1] \text{ and } \{\beta_i\}_{i\geq 0} \text{ with } \beta_{k+1} \geq \beta_k > L, \quad \forall k \geq 0.$$

But now the two sequences must satisfy another coupling condition:

$$\alpha_k^2 \beta_k \leq (\sum_{i=0}^{k} \alpha_i)\beta_{k-1}, \quad \forall k \geq 1. \tag{7.13}$$

We define also $A_k = \sum_{i=0}^{k} \alpha_i$ and $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.

Algorithm 25 Stochastic Fast Gradient Method (SFGM)

Compute $x_0 = \arg\min_{x \in Q} d(x)$
Let ξ_0 be a realization of the random variable X_0
Compute $G_{\delta,L}(x_0, \xi_0)$
Compute

$$y_0 = \arg\min_{x \in Q} \{\beta_0 d(x) + \alpha_0 \langle G_{\delta,L}(x_0, \xi_0), x - x_0 \rangle + h(x)\} \tag{7.14}$$

for k = 0 : ... **do**
    Compute

$$z_k = \arg\min_{x \in Q} \{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i, \xi_i), x - x_i \rangle + A_k h(x)\} \tag{7.15}$$

    Let

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k \tag{7.16}$$

    Let ξ_{k+1} be a realization of the random variable X_{k+1}
    Compute $G_{\delta,L}(x_{k+1}, \xi_{k+1})$
    Compute

$$\hat{x}_{k+1} = \arg\min_{x \in Q} \{\beta_k V(x, z_k) + \alpha_{k+1} \langle G_{\delta,L}(x_{k+1}, \xi_{k+1}), x - z_k \rangle + \alpha_{k+1} h(x)\} \tag{7.17}$$

    Let

$$y_{k+1} = \tau_k \hat{x}_{k+1} + (1 - \tau_k) y_k. \tag{7.18}$$

end for

This method is a generalization to the stochastic case of the Fast Gradient Method using only prox-function type subproblems introduced in [59] and described in subsection 2.4.5 of this thesis. It is based on the machinery of estimates functions (providing a more flexible method) and can be used easily with a non-Euclidean setup since it is based only on subproblems in terms of the prox-function $d(x)$.

In this work, we adapt the fast gradient method to the stochastic case, develop a new practical policy for the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ and prove that with this choice the method can reach the unimprovable rate of convergence $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ in unbiased stochastic smooth convex optimization.

The optimal rate $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}\right)$ has been obtained for the first time by Lan in [44] using an accelerated version of the Mirror Descent SA method with fixed stepsize based on the performed number of iterations. However, our method based on the estimates sequence principle does not assume the a priori knowledge of the number of iterations and does not assume the boundedness of the feasible set. Furthermore, our analysis consider also the composite case when we add to f an easy convex function $h(x)$ and the situation when the oracle is not only stochastic but also affected by a bias δ. We obtain a convergence rate of the form $O\left(\frac{LR^2}{k^2}+\frac{\sigma R}{\sqrt{k}}+k\delta\right)$. There is a phenomenon of errors accumulation with rate $\Theta(k\delta)$. It has been established in Chapter 4 that this is in fact unavoidable for any fast first-order method that reach the optimal dependence with respect to L in the convergence rate (i.e. $O\left(\frac{LR^2}{k^2}\right)$).

7.5.2 General convergence rate

Denote by $\Psi_k(x)=\beta_k d(x)+\sum_{i=0}^k \alpha_i[F_{\delta,L}(x_i,\xi_i)+\langle G_{\delta,L}(x_i,\xi_i),x-x_i\rangle+h(x)]$, our model of the objective function, $\Psi_k^*=\min_{x\in Q}\Psi_k(x)$ its minimal value on the feasible set and $\xi_{[k]}=(\xi_0,...,\xi_k)$ the history of the random process after k iterations.
Let us show that $\{y_k\}_{k\geq 0}$ and $\{\Psi_k(x)\}_{k\geq 0}$ define a sequence of estimate functions.

Lemma 7.7. *For all $k\geq 0$, we have:*

1.

$$A_k\phi(y_k)\leq \Psi_k^*+E_k \tag{7.19}$$

where $E_k=\sum_{i=0}^k A_i\delta+\sum_{i=0}^k \frac{A_i}{(\beta_i-L)}(\|G(x_i,\xi_i)-g_{\delta,L}(x_i)\|_E^*)^2$ $+\sum_{i=0}^k \alpha_i(f_{\delta,L}(x_i)-F_{\delta,L}(x_i,\xi_i))+\sum_{i=1}^k A_{i-1}\langle g_{\delta,L}(x_i)-G_{\delta,L}(x_i,\xi_i),x_i-y_{i-1}\rangle$.

2.

$$\Psi_k(x)\leq A_k\phi(x)+\beta_k d(x)+\overline{E}_k(x),\quad \forall x\in Q \tag{7.20}$$

where $\overline{E}_k(x)=\sum_{i=0}^k \alpha_i[F_{\delta,L}(x_i,\xi_i)-f_{\delta,L}(x_i)+\langle G_{\delta,L}(x_i,\xi_i)-g_{\delta,L}(x_i),x-x_i\rangle]$

Proof. 1. First, we want to prove by recurrence that the inequality $A_k\phi(y_k)\leq \Psi_k^*+E_k$ is satisfied for all $k\geq 0$.

• It is true for $k = 0$. Indeed:

$$\begin{aligned}
\Psi_0^* &\overset{(7.14)}{=} \beta_0 d(y_0) + \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)] \\
&\overset{(2.8)}{\geq} \frac{\beta_0}{2}\|y_0 - x_0\|_E^2 + \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)] \\
&\geq \alpha_0[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0) + \frac{\beta_0}{2}\|y_0 - x_0\|_E^2] \\
&= \alpha_0[f_0 + \langle g_0, y_0 - x_0\rangle + h(y_0) + \frac{L}{2}\|y_0 - x_0\|_E^2] \\
&\quad + \alpha_0[F_0 - f_0 + \langle G_0 - g_0, y_0 - x_0\rangle + \frac{\beta_0 - L}{2}\|y_0 - x_0\|_E^2] \\
&\overset{(7.2),(7.6)}{\geq} \alpha_0[f(y_0) + h(y_0) - \delta] + \alpha_0[F_0 - f_0] \\
&\quad - \frac{\alpha_0}{\beta_0 - L}(\|G_0 - g_0\|_E^*)^2.
\end{aligned}$$

• Assume that it is true for $k \geq 0$ i.e that we have $A_k\phi(y_k) \leq \Psi_k^* + E_k$. Let $gh(z_k) \in \partial h(z_k)$, by the optimality condition of the problem defining z_k:

$$\langle \beta_k \nabla d(z_k) + \sum_{i=0}^{k} G_i + A_k gh(z_k), x - z_k\rangle \geq 0, \quad \forall x \in Q.$$

Therefore as $\beta_{k+1} \geq \beta_k$:

$$\begin{aligned}
\Psi_{k+1}(x) &= \beta_{k+1} d(x) + \sum_{i=0}^{k+1} \alpha_i[F_i + \langle G_i, x - x_i\rangle] + A_{k+1}h(x) \\
&\geq \beta_k V(x, z_k) + \beta_k d(z_k) + \beta_k\langle \nabla d(z_k), x - z_k\rangle \\
&\quad + \sum_{i=0}^{k+1} \alpha_i[F_i + \langle G_i, x - x_i\rangle] + A_{k+1}h(x) \\
&\geq \beta_k V(x, z_k) + \beta_k d(z_k) + \sum_{i=0}^{k} \alpha_i[F_i + \langle G_i, z_k - x_i\rangle] \\
&\quad + A_{k+1}h(x) + \langle A_k gh(z_k), z_k - x\rangle \\
&\quad + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle].
\end{aligned}$$

But as $A_{k+1}h(x) + \langle A_k gh(z_k), z_k - x\rangle \geq A_k h(z_k) + \alpha_{k+1}h(x)$, we have:

$$\begin{aligned}
\Psi_{k+1}(x) &\geq \beta_k d(z_k) + \sum_{i=0}^{k} \alpha_i[F_i + \langle G_i, z_k - x_i\rangle + h(z_k)] \\
&\quad + \beta_k V(x, z_k) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
&\overset{(7.15)}{=} \Psi_k^* + \beta_k V(x, z_k) + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)].
\end{aligned}$$

On the other hand:

$$
\begin{aligned}
& \Psi_k^* + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
\geq\ & A_k\phi(y_k) - E_k + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)] \\
\overset{(7.2)}{\geq}\ & A_k[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] - E_k \\
& + \alpha_{k+1}[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle] + A_k h(y_k) + \alpha_{k+1}h(x) \\
=\ & A_{k+1}F_{k+1} + \langle G_{k+1}, A_k(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_k h(y_k) + \alpha_{k+1}h(x) \\
\overset{(7.16)}{=}\ & A_{k+1}F_{k+1} + \alpha_{k+1}\langle G_{k+1}, x - z_k\rangle - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_k h(y_k) + \alpha_{k+1}h(x).
\end{aligned}
$$

We obtain:

$$
\begin{aligned}
\Psi_{k+1}^* \quad \geq\ & A_{k+1}F_{k+1} + \min_{x\in Q}\{\beta_k V(x, z_k) + \alpha_{k+1}\langle G_{k+1}, x - z_k\rangle \\
& + \alpha_{k+1}h(x)\} - E_k + A_k h(y_k) \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
\overset{(7.17)}{=}\ & A_{k+1}F_{k+1} + \beta_k V(\hat{x}_{k+1}, z_k) + \alpha_{k+1}\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle \\
& + \alpha_{k+1}h(\hat{x}_{k+1}) - E_k + A_k h(y_k) \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
\overset{(2.10)}{\geq}\ & A_{k+1}[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle \\
& + \frac{\beta_k}{2A_{k+1}}\|\hat{x}_{k+1} - z_k\|_E^2] - E_k \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
& + A_{k+1}[\tau_k h(\hat{x}_{k+1}) + (1 - \tau_k)h(y_k)] \\
\overset{(7.13),(7.18)}{\geq}\ & A_{k+1}[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle \\
& + \frac{\beta_{k+1}\tau_k^2}{2}\|\hat{x}_{k+1} - z_k\|_E^2] - E_k + A_{k+1}h(y_{k+1}) \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle] \\
\overset{(7.16),(7.18)}{\geq}\ & A_{k+1}[F_{k+1} + \langle G_{k+1}, y_{k+1} - x_{k+1}\rangle \\
& + \frac{\beta_{k+1}}{2}\|y_{k+1} - x_{k+1}\|_E^2] - E_k + A_{k+1}h(y_{k+1}) \\
& + A_k[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle]
\end{aligned}
$$

and therefore:

$$\begin{aligned}
\Psi^*_{k+1} &= A_{k+1}[f_{k+1} + \langle g_{k+1}, y_{k+1} - x_{k+1}\rangle \\
&\quad + \frac{L}{2}\|y_{k+1} - x_{k+1}\|_E^2] - E_k + \alpha_{k+1}[F_{k+1} - f_{k+1}] \\
&\quad + A_k\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle \\
&\quad + A_{k+1}[\langle G_{k+1} - g_{k+1}, y_{k+1} - x_{k+1}\rangle \\
&\quad + \frac{\beta_{k+1} - L}{2}\|y_{k+1} - x_{k+1}\|_E^2] + A_{k+1}h(y_{k+1}) \\
&\overset{(7.2),(7.6)}{\geq} A_{k+1}(f(y_{k+1}) + h(y_{k+1}) - \delta) - E_k \\
&\quad + \alpha_{k+1}[F_{k+1} - f_{k+1}] + A_k\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle \\
&\quad - \frac{A_{k+1}}{\beta_{k+1} - L}(\|G_{k+1} - g_{k+1}\|_E^*)^2.
\end{aligned}$$

The inequality is therefore also satisfied for $k+1$ and we have proved our recurrence.

2. Now let us prove that (7.20) is satisfied for all $x \in Q$ and $k \geq 0$. Indeed:

$$\begin{aligned}
\Psi_k(x) &= \beta_k d(x) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x - x_i\rangle] + \sum_{i=0}^k \alpha_i[F_i - f_i] \\
&\quad + \sum_{i=0}^k \alpha_i[\langle G_i - g_i, x - x_i\rangle] + A_k h(x) \\
&\overset{(7.2)}{\leq} \beta_k d(x) + A_k(f(x) + h(x)) + \sum_{i=0}^k \alpha_i[F_i - f_i] \\
&\quad + \sum_{i=0}^k \alpha_i[\langle G_i - g_,x - x_i\rangle].
\end{aligned}$$

□

As we have proved that we are in the framework of estimate functions, we can now obtain directly the convergence rate for the SFGM:

Theorem 7.8. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then the sequence y_k generated by the Stochastic Fast Gradient Method, when applied to the composite function ϕ, satisfies*

$$\phi(y_k) - \phi^* \leq \frac{1}{A_k}\left(\beta_k d(x^*) + \sum_{i=0}^k A_i\delta + \sum_{i=0}^k \frac{A_i}{\beta_i - L}(\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_E^*)^2\right)$$

$$+ \frac{1}{A_k}(\sum_{i=1}^k A_{i-1}\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), y_{i-1} - x_i\rangle$$

$$+\sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle).$$

Taking the expectation with respect to $\xi_{[k]}$, the history of the random process, we obtain the following result:

Theorem 7.9. *Assume that the function f is endowed with a stochastic oracle with noise σ and bias δ. Then the Stochastic Fast Gradient Method, when applied to the composite function ϕ, exhibits on average the convergence rate*

$$E_{\xi_1\sim X_1,\ldots,\xi_k\sim X_k}[\phi(y_k) - \phi^*] \leq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i\delta}{A_k} + \frac{1}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}\sigma^2.$$

Proof. Same proof that for the Theorem 7.2 but now using the fact that y_{i-1} is also a deterministic function of $\xi_{[i-1]}$. □

7.5.3 Choice of the Coefficients

In the deterministic smooth case, the coefficients of the fast gradient method are chosen as $\beta_i = L$ and $\alpha_i = \frac{i+1}{2}$ for all $i \geq 0$.

If we keep these coefficients in the stochastic case, we cannot apply the Theorem 7.8 (that assumes $\beta_i > L$) but with an easy modification in the proof of this theorem, we can simply replace the term: $\sum_{i=0}^{k} \frac{A_i}{\beta_i - L} \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_*^2$ by $\sum_{i=0}^{k} A_i\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x_i - y_i\rangle$ in the upper-bound given by this theorem.
But as y_i depends on $G_{\delta,L}(x_i)$, we cannot say that $E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - y_i\rangle|\xi_{[i-1]}] = 0$ but only

$$\begin{aligned} & E[\langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - y_i\rangle|\xi_{[i-1]}] \\ \leq\ & \sqrt{E[(\|G_{\delta,L}(x_i) - g_{\delta,L}(x_i)\|_E^*)^2|\xi_{[i-1]}]}D \\ \leq\ & \sigma D \end{aligned}$$

where $D = \max_{x\in Q, y\in Q}\|x - y\|$ is the diameter of the feasible set. Therefore we have:

$$E[\phi(\hat{x}_k) - \phi^*] \leq \frac{2LR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta + \frac{1}{3}(k+3)D\sigma.$$

We see that with the classical choice of the coefficients, the effect of the stochastic noise σ does not decrease with the iterations like what we want to obtain. But in fact, it does not even stay constant like what we have obtained for the SPGM and SDGM with classical coefficients. Here the situation is even worse,

the effect of the noise is increasing with the number of iterations, there is a phenomenon of error accumulation. This higher sensitivity of the fast gradient method with respect to the noise has been already observed in Chapter 4 and in [75] when the error is deterministic. We have established that it is an intrinsic property of any fast first-order method with optimal convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$. In our case, it means that a dependence in the bias δ of the form $\Theta(k\delta)$ is unavoidable.

However, concerning the stochastic noise σ, the situation is better, we can modify the sequence of coefficients β_i in order to avoid this increasing dependence in σ in the convergence rate.

If we consider $\beta_i = CL$ with $C > 1$, we can apply Theorems 7.8, 7.9 and obtain:

$$E[\phi(\hat{y}_k) - \phi^*] \leq \frac{2CLR^2}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta + \frac{1}{3(C-1)L}(k+3)\sigma^2.$$

But we obtain the same kind of bad behavior with an accumulation of errors both for the stochastic part σ and the deterministic bias δ.

In this subsection, we want to develop a practical stepsizes rule for the stochastic fast gradient method which is not based on a a priori knowledge of the performed number of iterations and at the same times that can reach the convergence rate $\Theta(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta)$.
Consider the choice $\alpha_i = \frac{i+1}{a}$ and $\beta_i = L + b\frac{\sigma}{R}(i+2)^c$. Then we have $A_k = \sum_{i=0}^{k} \alpha_i = \frac{1}{2a}(k+1)(k+2)$ and the condition 7.13 becomes: $\frac{(k+1)^2}{a^2}(L + \frac{\sigma}{R}b(k+2)^c) \leq \frac{(k+1)(k+2)}{2a}(L + \frac{\sigma}{R}b(k+1)^c)$. A sufficient condition is to have:

1. $\frac{(k+1)^2}{a^2} \leq \frac{(k+1)(k+2)}{2a}$ for all $k \geq 0$ i.e. $a \geq 2$

2. $\frac{(k+1)^2}{a^2}\frac{\sigma}{R}b(k+2)^c \leq \frac{(k+1)(k+2)}{2a}\frac{\sigma}{R}b(k+1)^c$ for all $k \geq 0$ i.e. $a \geq 2^c$.

Assuming that $c \geq 1$, we choose $a = 2^c$. Then the condition $\alpha_k^2\beta_k \leq A_k\beta_{k-1}$ is satisfied, independently of the precise choice of b and c. With the choice of the sequences $\alpha_i = \frac{i+1}{2^c}$ and $\beta_i = L + \frac{\sigma}{R}b(i+2)^c$, we obtain $\frac{\beta_k R^2}{2A_k} =$

$\frac{2^c(L+\frac{\sigma}{R}b(k+2)^c)R^2}{(k+1)(k+2)}$, $\frac{\sum_{i=0}^k A_i}{A_k} = \frac{1}{3}(k+3)\delta$ and

$$\begin{aligned}\frac{1}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}\sigma^2 &= \frac{\sigma R}{(k+1)(k+2)b}\sum_{i=0}^{k}\frac{(i+1)}{(i+2)^{c-1}} \\ &\leq \frac{\sigma R}{(k+1)(k+2)b}\int_1 k+1(x+2)^{2-c}dx \\ &\leq \frac{\sigma R}{b(3-c)}\frac{(k+3)^{3-c}}{(k+1)(k+2)}.\end{aligned}$$

The bound given by Theorem 7.9 becomes as follows;

$$E(\phi(y_k) - \phi^*] \leq \frac{2^c LR^2}{(k+1)(k+2)} + \frac{2^c b\sigma R(k+2)^{c-1}}{(k+1)}$$

$$+\frac{\sigma R}{b(3-c)}\frac{(k+3)^{3-c}}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.$$

Now if we choose $c = 3/2$, the two terms depending on b and c are of order $\Theta(\frac{\sigma R}{k^{1/2}})$ and we obtain:

$$\begin{aligned}E_{\xi_1\sim X_1,\ldots\xi_k\sim X_k}[\phi(y_k) - \phi^*] &\leq \frac{2^{3/2}LR^2}{(k+1)(k+2)} \\ &\quad +\frac{(2^{3/2}b + \frac{2}{3b})(k+3)^{3/2}\sigma R}{(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.\end{aligned}$$

The optimal choice of b is $\frac{1}{2^{1/4}\sqrt{3}}$ and we obtain in this case the final result:

Theorem 7.10. *If the sequences $\{\alpha_i\}_{i\geq 0}$ and $\{\beta_i\}_{i\geq 0}$ are chosen in the following way: $\alpha_i = \frac{i+1}{2\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{1/4}\sqrt{3R}}(i+2)^{3/2}$ for all $i \geq 0$ then the sequence generated by the SFGM satisfies:*

$$\begin{aligned}E[\phi(y_k) - \phi^*] &\leq \frac{2^{3/2}LR^2}{(k+1)(k+2)} + \frac{2^{9/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)} + \frac{1}{3}(k+3)\delta \\ &= \Theta\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta\right).\end{aligned}$$

Remark 7.12. Due to the higher sensitivity of the FGM with respect to the stochastic noise σ, we need to increase the sequences of coefficients β_i at a fast rate $\Theta(L + \frac{\sigma}{R}i^{3/2})$ in order to decrease the stochastic noise at an optimal rate $O\left(\frac{\sigma R}{\sqrt{k}}\right)$. For the DGM which is more robust with respect to the errors, the increase of β_i can be limited to the rate $\Theta(L + \frac{\sigma}{R}i^{1/2})$.

7.6 Probability of large deviation

In the previous sections, we have obtained for different stochastic first-order methods, an upper bound on the expected value of the non-optimality gap $\phi(y_k) - \phi^*$. Now we want also to obtain an upper bound on the probability of large deviation for the same gap. The approach presented in this section is strongly linked with has been done in [52] for the mirror descent SA method in the nonsmooth stochastic case.
In this section, we need the following assumption:
Assumption A76

1. For all $x \in E$, the random variables X have the same distribution such that $X_0, ..., X_k$ can be seen as $i.i.d.$ random variables.

2. The stochastic approximate gradient $G_{\delta,L}(x,\xi)$ satisfies the condition $E_{\xi\sim X}\left[\exp\left(\frac{(\|G_{\delta,L}(x,\xi)-g_{\delta,L}(x)\|_E^*)^2}{\sigma^2}\right)\right] \leq \exp(1), \quad \forall x \in Q.$
Due to Jensen's inequality, this assumption is stronger that the assumption that we have done previously: $E_{\xi\sim X}[(\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_E^*)^2] \leq \sigma^2, \quad \forall x \in Q.$

3. The set Q is bounded with diameter $D = \max_{x\in Q, y\in Q} \|x - y\|_E$.

First of all, we establish two lemmas that will be useful in order to derive probability of large deviations for different first-order methods.

Lemma 7.11. *Let $\xi_0, ..., \xi_k$ be a sequence of realizations of the i.i.d. random variables $X_0, ..., X_k$ and let $\Delta_i = \Delta_i(\xi_{[i]})$ be a deterministic function of $\xi_{[i]}$ such that for all $i \geq 0$:*

$$E[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)|\xi_{[i-1]}] \leq \exp(1)$$

and $c_0, ..., c_k$ is a sequence of positive coefficients. Then we have for any $k \geq 0$ and any $\Omega \geq 0$:

$$Prob\left(\sum_{i=0}^{k} c_i\Delta_i^2 \geq (1+\Omega)\sum_{i=0}^{k} c_i\sigma^2\right) \leq \exp(-\Omega).$$

Proof. Using the convexity of the exponent and the linearity of the expectation, we

obtain:

$$\begin{aligned} & E\left[\exp\left(\frac{\sum_{i=0}^{k} c_i\Delta_i^2}{\sum_{i=0}^{k} c_i\sigma^2}\right)\right] \\ \leq & \frac{\sum_{i=0}^{k} c_i\sigma^2 E\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)\right]}{\sum_{i=0}^{k} c_i\sigma^2} \\ = & \frac{\sum_{i=0}^{k} c_i\sigma^2 E_{\xi_0\sim X_0,\ldots,\xi_i\sim X_i}\left[E_{\xi_i\sim X_i}\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)|\xi_{[i-1]}\right]\right]}{\sum_{i=0}^{k} c_i\sigma^2} \\ \leq & \exp(1). \end{aligned}$$

Therefore by the Markov inequality, for any $\tilde{\Omega} > 0$ we obtain:

$$\text{Prob}\left(\exp\left(\frac{\sum_{i=0}^{k} c_i\Delta_i^2}{\sum_{i=0}^{k} c_i\sigma^2}\right) \geq \tilde{\Omega}\right) \leq \frac{\exp(1)}{\tilde{\Omega}}.$$

Equivalently for any $\Omega \in \mathbb{R}$, we obtain:

$$\text{Prob}\left(\exp\left(\frac{\sum_{i=0}^{k} c_i\Delta_i^2}{\sum_{i=0}^{k} c_i\sigma^2}\right) \geq \exp(1+\Omega)\right) \leq \exp(-\Omega).$$

□

Lemma 7.12. *Let $\xi_0, ...\xi_0$ be a sequence of realizations of the i.i.d. random variables $X_0, ..., X_k$ and let Γ_k and η_k be deterministic functions of $\xi_{[k]}$ such that:*

1. $E[\Gamma_i|\xi_{[i-1]}] = 0$
2. $|\Gamma_i| \leq c_i\eta_i$ *where c_i is a positive deterministic constant*
3. $E[\exp\left(\frac{\eta_i^2}{\sigma^2}\right)|\xi_{[i-1]}] \leq \exp(1)$.

Then $Prob\left(\sum_{i=0}^{k}\Gamma_i \geq \sqrt{3}\sqrt{\Omega}\sigma\sqrt{\sum_{i=0}^{k} c_i^2}\right) \leq \exp(-\Omega)$ *for all $k \geq 0$ and all $\Omega \geq 0$.*

Proof. This result is a particular case of Lemma 2 in [45]. □

Now we are able using these two lemmas to establish easily probability of large deviation for the SDGM and the SFGM.

7.6.1 Probability of large deviation for SDGM

In the SDGM, the non-optimality gap $\phi(y_k) - \phi^*$ can be bounded by the sum of three terms (see Theorem 7.4) :

1. $H_1(k) = \frac{1}{A_k}\beta_k d(x^*) + \delta$
2. $H_2(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^{k} \frac{\alpha_i}{\beta_i - L}(\|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_E^*)^2$
3. $H_3(k, \xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$.

The first term is deterministic but the two others are random. Therefore in order to obtain a probability of large deviation for $\phi(y_k) - \phi^*$, a natural approach is to obtain probability of large deviation for $H_2(k, \xi_{[k]})$ and $H_3(k, \xi_{[k]})$ separately.

For $H_2(k, \xi_{[k]})$, using the Lemma 7.11 with $\Delta_i = \|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_E^*$ and $c_i = \frac{\alpha_i}{A_k(\beta_i - L)}$, we obtain that for any $k \geq 0$ and for any $\Omega \geq 0$:

$$Prob\left(H_2(k, \xi_{[k]}) \geq \frac{1+\Omega}{A_k}\sum_{i=0}^{k}\frac{\alpha_i}{\beta_i - L}\sigma^2\right) \leq \exp(-\Omega).$$

For $H_3(k, \xi_{[k]})$, using the Lemma 7.12 with $\Gamma_i = \frac{\alpha_i}{A_k}\langle G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$, $\eta_i = \|G_{\delta,L}(x_i, \xi_i) - g_{\delta,L}(x_i)\|_E^*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain that for any $k \geq 0$ and for any $\Omega \geq 0$:

$$Prob\left(H_3(k, \xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right) \leq \exp(-\Omega).$$

In conclusion, we obtain the following probability of large deviation for the SDGM:

Theorem 7.13. *If the assumption A76 is satisfied, then for all $k \geq 0$ and all $\Omega \geq 0$, the sequence generated by the SDGM satisfies*

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{\beta_k d(x^*)}{A_k} + \delta + \frac{(1+\Omega)}{A_k}\sum_{i=0}^{k}\frac{\alpha_i}{\beta_i - L}\sigma^2 + \frac{\sqrt{3\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right)$$
$$\leq 2\exp(-\Omega).$$

Using in particular the optimal coefficients policy $\alpha_i = \frac{1}{\sqrt{2}}$ and $\beta_i = L + \frac{2^{1/4}\sigma}{R}(i+1)^{1/2}$ for all $i \geq 0$, we obtain that for all $k \geq 0$ and all $\Omega \geq 0$:

$$Prob\left(\phi(y_k) - \phi^* \geq \Gamma_0(k) + \Gamma_1(k) + \Gamma_2(k) + \Gamma_3(k)\right) \leq 2\exp(-\Omega)$$

where $\Gamma_0(k) = \frac{\sqrt{2}LR^2}{2(k+1)}$, $\Gamma_1(k) = \delta$, $\Gamma_2(k) = \frac{2^{3/4}\sigma R}{\sqrt{k+1}}$ and $\Gamma_3(k) = \frac{\Omega\sigma R}{2^{1/4}\sqrt{k+1}} + \frac{\sqrt{3\Omega}D\sigma}{\sqrt{k+1}}$.

Remark 7.13. By Theorem 7.6, we have $E[\phi(y_k)-\phi^*] \leq \Gamma_0(k)+\Gamma_1(k)+\Gamma_2(k)$ and $\Gamma_3(k)$ represents therefore the deviation from the expected non-optimality gap.

Therefore a sufficient condition for ensuring $Prob(\phi(y_k) - \phi^* \geq \epsilon) \leq 1 - \gamma$ with $0 < \gamma < 1$, is to perform

$$k = \max\left(\frac{4LR^2}{\epsilon}, \frac{71\sigma^2R^2}{\epsilon^2}, \frac{18\sigma^2R^2}{\epsilon^2}\ln^2\left(\frac{2}{1-\gamma}\right), \frac{75\sigma^2D^2}{\epsilon^2}\ln\left(\frac{2}{1-\gamma}\right)\right)$$

iterations with $\delta \leq \frac{\epsilon}{5}$.

Remark 7.14. Exactly the same kind of analysis can be done for SPGM using Theorem 7.1 and Lemma 7.11 and 7.12. For this method, the probability of large deviation is given by :

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{V(x^*,x_0)}{\sum_{i=0}^{k-1}\gamma_i} + \delta + \frac{(1+\Omega)}{\sum_{i=0}^{k-1}\gamma_i}\sum_{i=0}^{k-1}\frac{\gamma_i}{\beta_i - L}\sigma^2 + \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{\sum_{i=0}^{k-1}\gamma_i}\sqrt{\sum_{i=0}^{k-1}\gamma_i^2}\right)$$
$$\leq 2\exp(-\Omega)$$

for all $k \geq 0$ and all $\Omega \geq 0$.

7.6.2 Probability of large deviation for the SFGM

In Theorem 7.8, we have proved that for the SFGM, the gap $\phi(y_k) - \phi^*$ can be bounded by the sum of four quantities:

1. $I_1(k) = \frac{1}{A_k}\left(\beta_k d(x^*) + \sum_{i=0}^{k} A_i\delta\right)$
2. $I_2(k,\xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}(\|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i,\xi_i)\|_E^*)^2$
3. $$\begin{aligned} I_3(k,\xi_{[k]}) &= \frac{1}{A_k}\sum_{i=1}^{k} A_{i-1}\langle G_{\delta,L}(x_i,\xi) - g_{\delta,L}(x_i), y_{i-1} - x_i\rangle \\ &= \frac{1}{A_k}\sum_{i=1}^{k}\alpha_{i-1}\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle \end{aligned}$$
4. $I_4(k,\xi_{[k]}) = \frac{1}{A_k}\sum_{i=0}^{k}\alpha_i\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$.

The first term $I_1(k)$ is deterministic but the three others are random.

For $I_2(k,\xi_{[k]})$, we use Lemma 7.11 with $\Delta_i = \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_E^*$ and $c_i = \frac{A_i}{A_k(\beta_i - L)}$, we obtain:

$$Prob\left(I_2(k,\xi_{[k]}) \geq \frac{1+\Omega}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}\sigma^2\right) \leq \exp(-\Omega)$$

for any $k \geq 0$ and for any $\Omega \geq 0$.

For $I_3(k,\xi_{[k]})$, using Lemma 7.12 (starting however the sum at $i=1$ instead of $i=0$) with $\Gamma_i = \frac{\alpha_{i-1}}{A_k}\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle$, $\eta_i = \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_E^*$ and $c_i = \frac{\alpha_{i-1}D}{A_k}$, we obtain:

$$Prob\left(I_3(k,\xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=1}^{k}\alpha_{i-1}^2}\right) \leq \exp(-\Omega)$$

for any $k \geq 1$ and for any $\Omega \geq 0$.

For $I_4(k,\xi_{[k]})$, using Lemma 7.12 with $\Gamma_i = \frac{\alpha_i}{A_k}\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle$, $\eta_i = \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_E^*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain:

$$Prob\left(I_4(k,\xi_{[k]}) \geq \frac{\sqrt{3}\sqrt{\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right) \leq \exp(-\Omega)$$

for any $k \geq 0$ and for any $\Omega \geq 0$.

In conclusion, we obtain the following probability of large deviation for the SFGM:

Theorem 7.14. *Assume that assumption A76 is satisfied, then for all $k \geq 0$ and all $\Omega \geq 0$, the sequence generated bv the SFGM satisfies:*

$$Prob\left(\phi(y_k) - \phi^* \geq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i}{A_k}\delta + \frac{(1+\Omega)}{A_k}\sum_{i=0}^{k}\frac{A_i}{\beta_i - L}\sigma^2 + \frac{2\sqrt{3\Omega}D\sigma}{A_k}\sqrt{\sum_{i=0}^{k}\alpha_i^2}\right)$$

$$\leq 3\exp(-\Omega).$$

Using in particular the optimal coefficients policy i.e. $\alpha_i = \frac{i+1}{2\sqrt{2}}$ and $\beta_i = L + \frac{\sigma}{2^{1/4}\sqrt{3}R}(i+2)^{3/2}$ for all $i \geq 0$, we obtain:

$$\text{Prob } (\phi(y_k) - \phi^* > \Lambda_0(k) + \Lambda_1(k) + \Lambda_2(k) + \Lambda_3(k)) \leq 3\exp(-\Omega)$$

where $\Lambda_0(k) = \frac{2^{3/2}LR^2}{(k+1)(k+2)}$, $\Lambda_1(k) = \frac{k+3}{3}\delta$, $\Lambda_2(k) = \frac{2^{9/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)}$ and $\Lambda_3(k) = \frac{2^{5/4}\Omega\sigma R}{\sqrt{3}}\frac{(k+3)^{3/2}}{(k+1)(k+2)} + \frac{2\sqrt{\Omega}\sigma D}{\sqrt{3}}\sqrt{\frac{2k+3}{(k+1)(k+2)}}$.

Remark 7.15. By Theorem 7.10, we have $E[\phi(y_k) - \phi^*] \leq \Lambda_0(k) + \Lambda_1(k) + \Lambda_2(k)$ and $\Lambda_3(k)$ represents therefore the deviation from the expected non-optimality gap.

Therefore a sufficient condition for ensuring $Prob(\phi(y_k) - \phi^* \geq \epsilon) \leq 1 - \gamma$ with $0 < \gamma < 1$, is to perform

$$k = \max\left(4\sqrt{\frac{LR^2}{\epsilon}}, \frac{336\sigma^2R^2}{\epsilon^2}, \frac{84\sigma^2R^2}{\epsilon^2}\ln^2\left(\frac{3}{1-\gamma}\right), \frac{17\sigma^2D^2}{\epsilon^2}\ln\left(\frac{3}{1-\gamma}\right)\right)$$

with $\delta \leq \frac{\epsilon}{5}$.

7.7 Postoptimization: Accuracy certificate

After running k iterations of one of the stochastic first-order methods, we obtain a feasible point $y_k \in Q$ for the optimization problem (7.1).
We have obtained in the previous sections theoretical guarantees for the expected non optimality gap $\phi(y_k) - \phi^*$ and for the probability of large deviations of this gap from its expected value. However, we could be also interested in estimating the actual value of $\phi(y_k) - \phi^*$ since, in practice, the quality of y_k can be better that what is guaranteed by worst-case oriented theoretical bounds. If we want to estimate $\phi(y_k) - \phi^*$:

1. We need to compute $\phi(y_k) = f(y_k) + h(y_k)$ or at least a stochastic estimate of $\phi(y_k)$
2. We need to compute a lower bound on ϕ^* or at least a random number Φ^* which is on average (and with small probability of large deviation) a lower bound on ϕ^*.

In the deterministic case:

1. We can compute $\phi(y_k) = f(y_k) + h(y_k)$ using the exact oracle
2. we can obtain a lower bound on ϕ^*, minimizing on Q, the sum of h with the linearization of f at y_k:

$$\phi^* \geq \min_{x \in Q}\{f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + h(x)\}.$$

In the stochastic case, $f(y_k)$ and $\nabla f(y_k)$ are typically unavailable (or to costly to compute) and we will try to use accurate estimates of these quantities using our stochastic oracle. We proceed as follow:

1. We generate N independent samples $\eta_1, ..., \eta_N$ from the random variable Y_k
2. We compute $F_{\delta,L}(y_k, \eta_1), ..., F_{\delta,L}(y_k, \eta_N)$ and $G_{\delta,L}(y_k, \eta_1), ..., G_{\delta,L}(y_k, \eta_N)$ using the stochastic oracle
3. In order to reduce the noise, we construct better estimates of $f(y_k)$ and $\nabla f(y_k)$ using averaging:

$$\begin{aligned} F_{\delta,L}(y_k, \eta_1, ..., \eta_n) &= \frac{1}{N}\sum_{i=1}^{N} F_{\delta,L}(y_k, \eta_i) \text{ and} \\ G_{\delta,L}(y_k, \eta_1, ..., \eta_n) &= \frac{1}{N}\sum_{i=1}^{N} G_{\delta,L}(y_k, \eta_i). \end{aligned}$$

In this section, we make the following assumption:

Assumption A77

1. For all $x \in E$, the random variables X have the same distribution such that $X_0, ..., X_k$ can be seen as $i.i.d.$ random variables.
2. $E_{\xi\sim X}\{\exp\left(\frac{|F_{\delta,L}(x,\xi)-f_{\delta,L}(x)|^2}{\sigma_F^2}\right)\} \leq \exp(1)$
3. $E_{\xi\sim X}\{\exp\left(\frac{(\|G_{\delta,L}(x,\xi)-f_{\delta,L}(x)\|_E^*)^2}{\sigma_G^2}\right)\} \leq \exp(1)$
4. We have a zero-order oracle for the function h that can compute $h(x)$ for all $x \in Q$.
5. The set Q is bounded with diameter $D = \max_{x\in Q, y\in Q} \|x - y\|_E$.

Now we can obtain:

• **A good random estimate of** $\phi(y_k)$: $F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k)$
Indeed, we have:

$$\begin{aligned} \phi(y_k) - \delta &\leq E_{\eta_1\sim Y_k, ...\eta_N\sim Y_k}[F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k)] \\ &= f_{\delta,L}(y_k) + h(y_k) \leq \phi(y_k) \end{aligned}$$

and if we increase the number of samples N, we decrease the probability of deviation of $F_{\delta,L}(y_k, \eta_1, ..., \eta_n)$ from his expected value $f_{\delta,L}(y_k)$:
$Prob\left(|F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k)| \geq K\right) \leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}K}{\sqrt{2}\sigma_F} - 1\right)^2\right)$ (using the Theorem 2.1 (ii) in [35]) and therefore:

$$Prob\left(|F_{\delta,L}(y_k, \eta_1, ..., \eta_N) + h(y_k) - \phi(y_k)| \geq K + \delta\right)$$

$$\leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}K}{\sqrt{2}\sigma_F}-1\right)^2\right).$$

• **An approximate lower bound for ϕ^*:**

$$\Phi^* = \min_{x\in Q}\{F_{\delta,L}(y_k,\eta_1,...,\eta_N) + \langle G_{\delta,L}(y_k,\eta_1,...,\eta_N), x-y_k\rangle + h(x)\}$$

which on average, provides us with a lower bound on ϕ^*. The probability of deviation of Φ^* from being really a lower bound on ϕ^* decreases with the size of the sample. Indeed, we have:

Theorem 7.15. *For all $\beta \geq 0$:*

$$Prob(\phi^* \geq \Phi^* - \beta)$$

$$\geq 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F}-1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G}-\sqrt{\kappa}\right)^2\right)\right)$$

where κ is the constant of regularity of $(E,\|.\|)$ (see [35]).

Proof. Applying the Theorem 2.1 (ii) in [35] to $F_{\delta,L}(y_k,\eta_1,...,\eta_N)$, we obtain for all $\beta \geq 0$:

$$Prob\left(|F_{\delta,L}(y_k,\eta_1,...,\eta_N) - f_{\delta,L}(y_k)| \geq \frac{\beta}{2}\right) \leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F}-1\right)^2\right).$$

Applying the same theorem to $G_{\delta,L}(y_k,\eta_1,...,\eta_N)$, we obtain for all $\beta \geq 0$:

$$Prob\left(\|G_{\delta,L}(y_k,\eta_1,...,\eta_N) - g_{\delta,L}(y_k)\|_E^* \geq \frac{\beta}{2}\right)$$

$$\leq \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_G}-\sqrt{\kappa}\right)^2\right).$$

Now as

$$f(x) \geq f_{\delta,L}(y_k) + \langle g_{\delta,L}(y_k), x-y_k\rangle, \quad \forall x \in Q$$

we have:

$$\begin{aligned}
& Prob(\exists x \in Q : \phi(x) \leq F_{\delta,L}(y_k, \eta_1, ..., \eta_N) \\
& +\langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle + h(x) - \beta) \\
= \; & Prob(\exists x \in Q : f(x) \leq F_{\delta,L}(y_k, \eta_1, ..., \eta_N) \\
& +\langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle - \beta) \\
\leq \; & Prob(\exists x \in Q : F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k) \\
& +\langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N) - g_{\delta,L}(y_k), x - y_k\rangle \geq \beta) \\
\leq \; & \max(Prob\left(F_{\delta,L}(y_k, \eta_1, ..., \eta_N) - f_{\delta,L}(y_k) \geq \frac{\beta}{2}\right), \\
& Prob\left(\|G_{\delta,L}(y_k, \eta_1, ..., \eta_N) - g_{\delta,L}(y_k)\|_E^* \geq \frac{\beta}{2D}\right)) \\
\leq \; & \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right),
\end{aligned}$$

and therefore:

$$\begin{aligned}
& Prob\left(\forall x \in Q : \phi(x) \geq F_{\delta,L}(y_k, \eta_1, .., \eta_N) + \langle G_{\delta,L}(y_k, \eta_1, .., \eta_N), x - y_k\rangle + h(x) - \beta\right) \\
& \geq 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right).
\end{aligned}$$

In particular, we have:

$$\begin{aligned}
& Prob\left(\phi^* \geq \Phi^* - \beta\right) \\
= \; & Prob(\min_{x \in Q} \phi(x) \geq \min_{x \in Q}\{F_{\delta,L}(y_k, \eta_1, ..., \eta_N) \\
& +\langle G_{\delta,L}(y_k, \eta_1, ..., \eta_N), x - y_k\rangle + h(x)\} - \beta) \\
\geq \; & 1 - \max\left(\exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}\sigma_F} - 1\right)^2\right), \exp\left(-\frac{1}{3}\left(\frac{\sqrt{N}\beta}{2\sqrt{2}D\sigma_G} - \sqrt{\kappa}\right)^2\right)\right).
\end{aligned}$$

□

Remark 7.16. When $2 \leq p \leq +\infty$, the constant of regularity of $(\mathbb{R}^n, \|.\|_p)$ satisfies:

$$\kappa \leq \min(p - 1, 2\ln(n)).$$

The regularity constants of various other normed spaces can be found in [35].

7.8 Numerical Experiments: Quadratic Problem with Stochastic noise

In this section, we want to test the methods developed in this chapter (and to compare them with existing methods) on convex quadratic problems over the

simplex:

$$f^* = \min_{x \in \Delta_n} f(x) = \frac{1}{2} x^T A x \tag{7.21}$$

where $A \succeq 0$ and the l_1 setup is used.

Remark 7.17. As SPGM and SDGM share the same theoretical behavior and as the numerical results obtained using both methods are comparable, we do not consider in this section SPGM but only the methods that are really new in the stochastic context i.e. SDGM and SFGM.

In the exact case i.e. when the exact gradient $\nabla f(x) = Ax$ is available, the Fast Gradient Method (used with exact gradients and constant coefficients $\beta_i = L = \|A\|_\infty$) outperforms significantly the Dual Gradient Method (used with exact gradient and constant coefficients $\beta_i = L = \|A\|_\infty$). Performing 10 000 iterations, we obtain for $f(y_k) - f^*$:

Num. Iter.	10	100	1000	10000
DGM	0.478796	0.329690	0.0720594	0.0066759
FGM	0.427691	0.0233784	3.6576e-4	8.3417e-6

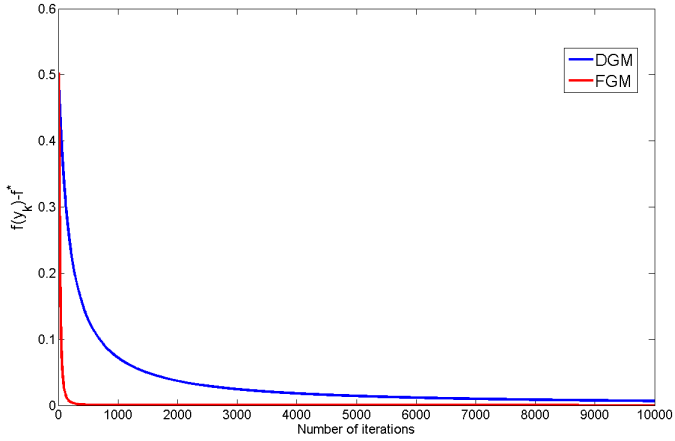


Figure 23: Comparison between DGM and FGM in the exact case ($\sigma = 0$).

This result is completely predicted by theory, the FGM exhibits a convergence rate of the form $\Theta\left(\frac{LR^2}{k^2}\right)$, significantly better than $\Theta\left(\frac{LR^2}{k}\right)$ for the DGM.

Now assume that we have only access to a stochastic gradient $G_{\delta,L}(x,\xi) = Ax + \xi$ where ξ is a stochastic noise (with normal distribution) such that $E[\xi] = 0$ and $E[\|\xi\|_*^2] \leq \sigma^2$. We consider first a reasonable noise level $\sigma = 1$ (corresponding to 1 % of the Lipschitz-constant of the gradient). We can try

to apply the SDGM and the SFGM with constant coefficients $\beta_i = L$ like what we do in the exact case. This choice is not recommended by the theory since the SDGM exhibits in this case a rate $\Theta\left(\frac{LR^2}{k} + \sigma D\right)$ and the SFGM $\Theta\left(\frac{LR^2}{k^2} + kD\sigma\right)$. Performing 10000 iterations, we obtain:

Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.481479	0.335265	0.0728529	0.00698266
SFGM (C=0)	0.428563	0.0385569	0.399419	0.881574

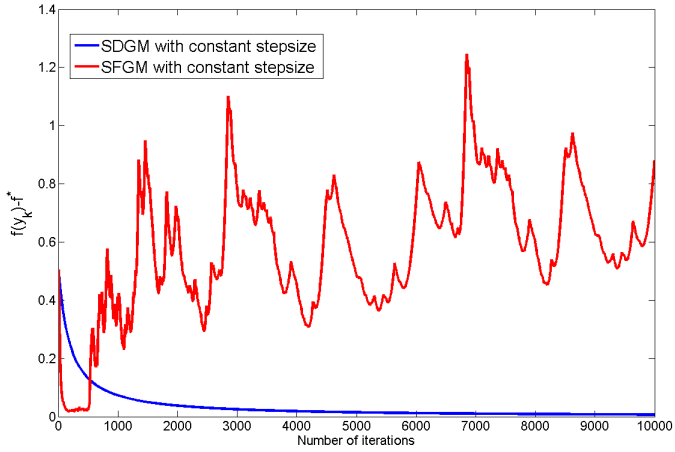


Figure 24: Comparison between SDGM and SFGM used with constant stepsize ($\sigma = 1$)

SDGM exhibits here a slow but convergent behavior. However, we see that SFGM is unstable and suffers from accumulation of errors. This bad behavior of the SFGM when used with constant stepsize ($\gamma_i = \frac{1}{L}$) and a stochastic oracle has been predicted by theory. The SDGM is slow but more robust to the errors, the method is still convergent even with this aggressive constant stepsize policy.

In order to avoid this sensitivity to the stochastic noise σ, we use now the decreasing stepsize policies developed in this chapter i.e. the increasing sequence of coefficients: $\beta_i = L + \frac{2^{1/4}C\sigma}{R}(i+1)^{1/2}$ for the SDGM and $\beta_i = L + \frac{C\sigma}{2^{1/4}\sqrt{3}R}(i+2)^{3/2}$ for the SFGM. When $C = 0$, we retrieve the constant stepsize policy and $C = 1$ corresponds to the theoretical optimal choice.
With the theoretical optimal choice C=1, we obtain:

Num. Iter.	10	100	1000	10000
SDGM (C=1)	0.481753	0.339013	0.0786420	0.00822247
SFGM (C=1)	0.431472	0.0531080	0.00491995	7.851197e-4

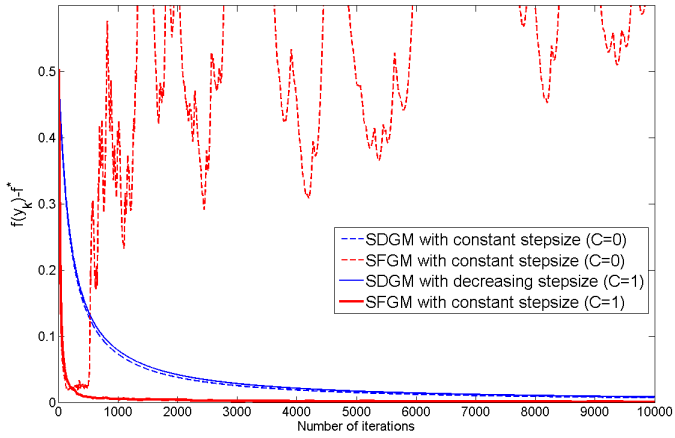


Figure 25: Comparison between SDGM and SFGM used with constant and decreasing stepsize ($\sigma = 1$)

The SFGM retrieves his good behavior, the method is significantly faster than the SDGM and can decrease now the effect of the oracle noise (instead of increasing it with constant stepsizes). We see here clearly the importance of using decreasing stepsizes in the stochastic case (at least for the fast-gradient method). For the SDGM, for this level of noise, a decreasing sequence of stepsize seems not necessary and slow down a little bit the convergence.

We can also compare our methods (SDGM and SFGM) with the methods developed by Lan in [44]:

- ◇ The Modified Mirror Descent SA (MMDSA) method with convergence rate $\Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ (like what we obtain for the SDGM when used with C=1)
- ◇ The Accelerated SA (AC-SA) method with convergence rate $\Theta\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ (like what we obtain for the SFGM when used with C=1).

An important property of the methods developed by Lan is the fact that they are based on the a priori knowledge of the performed number of iterations N. The goal of these methods is to reach a good accuracy after N iterations, not

for intermediate $0 < k < N$. Performing 10000 iterations of our two methods and the two methods developed by Lan, we obtain:

Num. Iter.	10	100	1000	10000
SDGM (C=1)	0.481753	0.339013	0.0786420	0.00822247
SFGM (C=1)	0.431472	0.0531080	0.00491995	7.851197e-4
MMDSA	0.491474	0.376019	0.0986267	0.0100789
AC-SA	0.508434	0.503937	0.249861	0.00365878

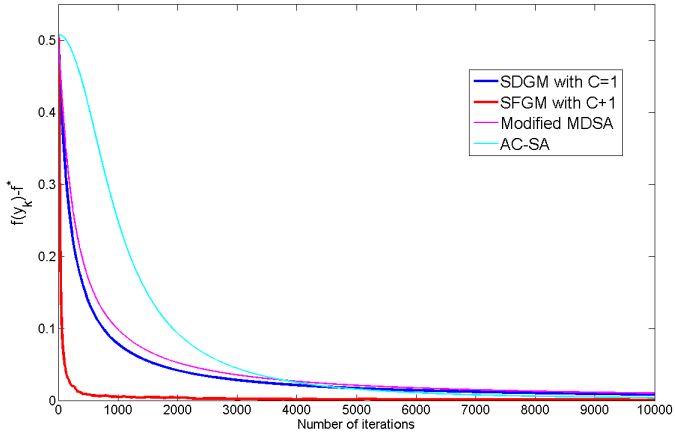


Figure 26: Comparison between SDGM, SFGM, Modified MDSA and AC-SA ($\sigma = 1$).

For the gradient-type methods (i.e. the SDGM and the MMDSA method), the two methods exhibits the same kind of behavior with however a faster convergence for our SDGM.
For the fast-gradient-type methods (i.e. the SFGM and the AC-SA method), the AC-SA is only efficient if we a total of N iterations, not for an intermediate number of iterations, whereas the SFGM is fast everywhere. We see here clearly the advantage of methods that are not based on a fixed number of iterations.

In conclusion, when the stochastic noise is reasonable (here 1 $\%$ of the Lipschitz-constant of the gradient), the SFGM with decreasing stepsize seems to be the method of choice. This method is fast (compared to SDGM and MMDSA method), is not sensitive to the oracle error (compared to SFGM with constant stepsize) and is flexible, does not need to perform exactly an a priori fixed number of iterations (which is the case for the AC-SA method).

We consider now the situation when the noise σ is significantly more important: $\sigma = 10$. First, we compare the SDGM with the SFGM, both using constant or decreasing stepsizes:

Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.526044	0.467577	0.155076	0.030702
SDGM (C=1)	0.523209	0.463160	0.1751157	0.0419122
SFGM (C=0)	0.48812	0.5741252	0.446167	0.975812
SFGM (C=1)	0.462503	0.292540	0.097984	0.026385

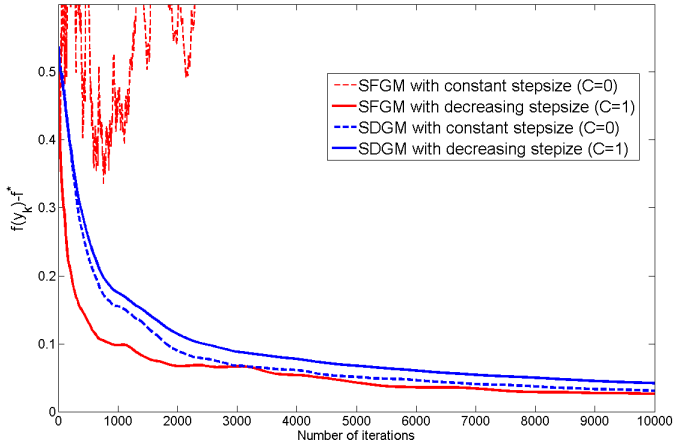


Figure 27: Comparison between SDGM and SFGM used with constant and decreasing stepsize ($\sigma = 10$)

We observe that:

- ◇ The SFGM must be used with decreasing stepsizes in order to avoid a bad accumulation errors. This phenomenon has been already observed for $\sigma = 1$.
- ◇ The SFGM with decreasing stepsize is a little bit faster than the SDGM with decreasing stepsize. However the advantage of the SFGM is significantly reduced compare to the case $\sigma = 1$. This is natural, when the noise is large, the advantage of a convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ over $O\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}}\right)$ becomes negligible, the dominant term in the convergence rate becomes quickly the bad term coming from the noise.
- ◇ The SDGM can be used with constant stepsize and this more aggressive choice gives a faster convergence. It seems that the robustness of the

SDGM (more important than expected by the theory) is sufficient in order to avoid a decreasing stepsize even when $\sigma = 10$. The worst-case oriented decreasing stepsize policy seems to slow down the method unnecessarily on this numerical example.

Now we can compare also our methods with the methods developed by Lan on this noisy example:

Num. Iter.	10	100	1000	10000
SDGM (C=0)	0.526044	0.467577	0.155076	0.030702
SDGM (C=1)	0.523209	0.463160	0.1751157	0.0419122
SFGM (C=0)	0.48812	0.5741252	0.446167	0.975812
SFGM (C=1)	0.462503	0.292540	0.097984	0.026385
MMDSA	0.494363	0.4463241	0.1633532	0.034726
AC-SA	0.508496	0.508166	0.4631923	0.0593871

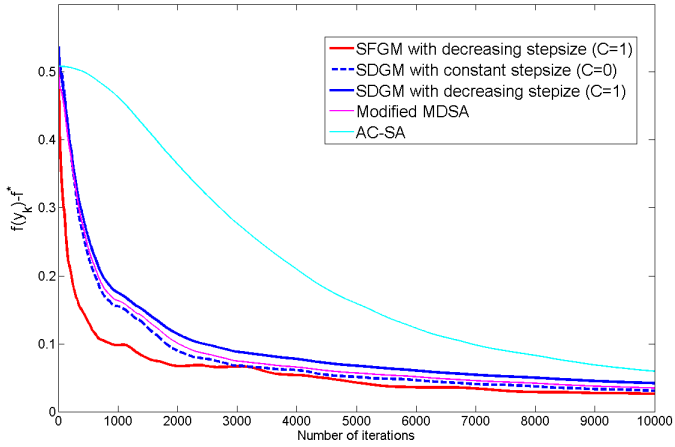


Figure 28: Comparison between SDGM, SFGM, Modified MDSA and AC-SA ($\sigma = 10$).

We observe that:

- ◇ The AC-SA method performs badly on this example. This method is very slow at the beginning (the method being designed only to reach a good accuracy after the fixed number of iterations N) and even ater the N iterations, the obtained solution is not so accurate. The SFGM with decreasing stepsize that share the same convergence rate $O\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}}\right)$ is clearly a better choice.

- ◇ The MMDSA method of Lan performs well on this noisy example but the best choice for a gradient type method seems to be the SDGM with aggressive constant stepsize.

Chapter 8

Conclusion

In this thesis, we have extended the scope of the first-order methods of smooth convex optimization in three different directions by allowing inexactness in the first-order information, lack of smoothness in the objective function and the presence of linear constraints.

In this chapter, for each of these situations, we summarize in the first section the main contributions obtained in this thesis and discuss in the second section potential directions of further research and possible extensions of our results. In addition to this conclusion chapter, the reader can find afterwards, a table of methods that emphasizes the new schemes that have been developed in this thesis.

8.1 Extended Summary

8.1.1 First-order methods with inexact first-order information (Chapters 4, 5, 6 and 7)

Two kinds of objective functions (smooth convex or smooth strongly convex) and two kinds of oracle errors (deterministic or stochastic) have been considered leading to four different possible situations:

	Deterministic oracle	Stochastic oracle
Smooth convex function	Chapter 4 and 6	Chapter 7
Smooth Strongly convex function	Chapter 5	Further Research

• Smooth convex function with deterministic inexact oracle (Chapters 4 and 6)

When exact first-order information is available, it is well-known that the Fast Gradient Method (FGM) outperforms the classical gradient methods (i.e. the Primal and Dual Gradient Methods) and is the first-order method of choice for solving smooth convex optimization problems.

One of the main messages of this thesis is that this clear superiority of the FGM is no longer true in the presence of inexactness in the first-order information. The situation is more balanced: the Primal and Dual Gradient Methods (PGM/DGM) are slow but robust methods while the Fast Gradient Method (FGM) is fast but sensitive to oracle errors. More precisely, we have obtained that PGM and DGM exhibit a slow convergence rate proportional to $\frac{1}{k}$ (where k denotes the iteration counter) but without accumulation of errors, the total effect of the errors after k iterations is simply equal to the individual error δ of each first-order information. On the other hand, FGM is faster with convergence rate proportional to $\frac{1}{k^2}$ but suffers from accumulation of errors at a linear rate $\Theta(k\delta)$.

	PGM/DGM	FGM.
Convergence Rate $f(y_k) - f^* \leq$	$\Theta(\frac{LR^2}{k}) + \delta$	$\Theta(\frac{LR^2}{k^2}) + \Theta(k\delta)$

In terms of complexity, it means that PGM (and DGM) can reach any target accuracy $\epsilon > \delta$ but needs a large number of iterations, proportional to $\frac{1}{\epsilon}$. The FGM has a better complexity proportional to $\sqrt{\frac{1}{\epsilon}}$ but is only able to reach worse target accuracies $\epsilon \geq \epsilon^*_{FGM} = \Theta(L^{1/3}R^{2/3}\delta^{2/3}) > \delta$.

	PGM/DGM	FGM.
Complexity	$\Theta(\frac{LR^2}{\epsilon})$	$\Theta(\sqrt{\frac{LR^2}{\epsilon}})$
Best Reachable Accuracy	δ	$\Theta(L^{1/3}R^{2/3}\delta^{2/3})$

These two very different behaviors lead to a natural desire to develop a new method sharing the best of the PGM/DGM and of the FGM i.e. a method which is as fast as the FGM (i.e. with a convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$ in the exact case) and as robust with respect to the oracle error as the PGM/DGM (i.e. without accumulation of error). In term of complexity, it would mean to obtain a

method that can reach any target accuracy $\epsilon \geq \delta$ in only $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations.

Unfortunately, this objective is too ambitious. We have proved that within our specific model of inexactness, accumulation of errors is not specific to the FGM itself but it is an intrinsic property of any fast first-order method. One of the main messages of this thesis is certainly the intrinsic link that exists between speed of convergence and sensitivity to errors for a first-order method: the faster a method is, the higher its sensitivity with respect to error must be. More precisely, a method with convergence rate proportional to $\frac{1}{k^p}$ $(1 \leq p \leq 2)$ must suffer from error accumulation with a rate at least proportional to $k^{p-1}\delta$:

$$f(x_k) - f^* \leq \Theta\left(\frac{LR^2}{k^p}\right) + \Theta(k^q\delta) \Rightarrow q \geq p - 1.$$

Nevertheless, between the two extreme choices of the robust but (very) slow PGM/DGM and the fast but (highly) sensitive FGM, it could be preferable to use a method with intermediate speed and intermediate sensitivity to errors. Such kind of intermediate behavior cannot be obtained by a simple stepsize modification in the existing methods or by a simple combination of them. Instead, we have developed a novel general family of methods, the Intermediate Gradient Method (IGM) that provides us with a large degree of freedom in the choice of two sequences of coefficients that are used in the scheme and characterize its convergence rate. The IGM allows us to recover the DGM and FGM as special cases but also to generate methods with intermediate behaviors.

With a switching policy for the coefficients, the IGM can be seen as a smart switching between FGM and DGM and is able to reach target accuracies, unreachable by the FGM (i.e. such that $\epsilon < \epsilon^*_{FGM}$) in a significantly smaller number of iterations compared to what is needed using the PGM/DGM.

	PGM/DGM	IGM with switching policy	FGM
Optimal when	$\epsilon \leq 2\delta$	$\epsilon_{MIN} = 2\delta < \epsilon < \epsilon_{MAX} = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$	$\epsilon_{MAX} \leq \epsilon$
Complexity	$\Theta\left(\frac{LR^2}{\epsilon}\right)$	$\Theta\left(\frac{LR^2\delta^2}{\epsilon^2}\right)$	$\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$

With another choice of the coefficients, we are able to generate, for any $1 \leq p \leq 2$, methods with convergence rate proportional to $\frac{1}{k^p}$ and corresponding optimal rate of error accumulation of order $k^{p-1}\delta$, exhibiting every possible trade-off between fastness of the method and robustness to errors.

	PGM/DGM	IGM with power policy	FGM.
Conv. Rate $f(y_k) - f^* \leq$	$\Theta\left(\frac{LR^2}{k}\right)+\delta$	$\Theta\left(\frac{LR^2}{k^p}\right) + \Theta(k^{p-1}\delta)$	$\Theta\left(\frac{LR^2}{k^2}\right) + \Theta(k\delta)$

With these intermediate gradient methods, we close in some sense the gap that existed previously between two extreme behaviors: the fast but sensitive FGM and the slow but robust PGM/DGM.

• Smooth strongly convex function with deterministic inexact oracle (Chapter 5)

Adding a strong convexity assumption on the objective function and still assuming a deterministic inexact oracle, the same kind of analysis can be done. We introduce the notion of (δ, L, μ)-oracle that can be seen as an extension of the (δ, L)-oracle, taking into account strong convexity.

As in the smooth convex case, we obtain that the simple gradient methods (PGM/DGM) can be seen as robust but slower, whereas the FGM is faster but more sensitive to oracle errors. However, strong convexity leads to much faster convergence rates for every method (linear instead of sublinear convergence rates) and to a reduced sensitivity with respect to oracle errors (bounded instead of unbounded accumulation of errors). The central quantity is now the condition number $\frac{L}{\mu}$ of the smooth strongly convex objective function.

The PGM/DGM exhibits a convergence rate proportional to $\exp(-k\frac{\mu}{L})$ and no accumulation of errors whereas the FGM exhibits a faster convergence rate proportional to $\exp(-k\sqrt{\frac{\mu}{L}})$ and an accumulation of errors that can be bounded by a constant proportional to $\sqrt{\frac{L}{\mu}}\delta$ (instead of a linearly growing accumulation of errors without strong convexity). FGM is less sensitive with respect to oracle errors than without strong convexity, but still more sensitive than the PGM/DGM.

	PGM/DGM	FGM.
Conv. Rate $f(y_k) - f^* \leq$	$\Theta\left(LR^2\exp(-k\frac{\mu}{L})\right)+\delta$	$\Theta\left(LR^2\exp(-k\sqrt{\frac{\mu}{L}})\right)+$ $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$
Complexity	$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)$	$\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)$
Best Reach. Acc.	δ	$\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$

Here also, we have established an intrinsic link between speed of convergence (more precisely the dependence in the condition number of the convergence rate) and sensitivity with respect to errors (more precisely the dependence in the condition number of the accumulation of errors). This result shows that there is no hope to obtain a method as fast as the FGM and as robust as the PGM/DGM but open the door to the development of new methods with intermediate behaviors.

• Smooth convex functions with stochastic inexact oracle (Chapter 7)

We have also considered the situation of a stochastic oracle, where the first-order information suffers from a stochastic noise σ. We have studied the behavior of our three first-order methods when used in this stochastic context, first without any modification of the methods, second with specially designed stepsizes taking into account the stochastic noise.

When used without modification, i.e. with constant stepsizes, the behaviors of the PGM/DGM/FGM are similar to what we have obtained in the deterministic case, the stochastic noise (multiplied by the diameter of the feasible set) replacing simply the deterministic error and the expected convergence rate replacing the deterministic one.

However, the stochastic case is in some sense more favorable compared to the deterministic case (at least when the oracle is unbiased, otherwise the bias plays the role of a deterministic error). Taking into account the stochastic nature of the first-order information, it is possible to improve the performance of these methods.

Whereas we cannot expect to obtain an accuracy on the objective function better than δ in the deterministic case, it is possible to expect a decrease of the stochastic noise effect up to 0 with a rate proportional to $\frac{1}{\sqrt{k}}$. Furthermore, such unimprovable decrease of the stochastic noise effect can be obtained simply by using well-chosen decreasing stepsizes in the PGM/DGM/FGM (that become with these modifications respectively the SPGM, the SDGM and the SFGM). In particular, the faster convergence rate of the SFGM does not prevent us from decreasing the stochastic noise at the same rate as for the SPGM/SDGM.

	SPGM/SDGM	SFGM.
Conv. Rate $E[f(y_k) - f^*]$	$\leq \Theta\left(\frac{LR^2}{k}\right) + \Theta\left(\frac{\sigma R}{\sqrt{k}}\right)$	$\leq \Theta\left(\frac{LR^2}{k^2}\right) + \Theta\left(\frac{\sigma R}{\sqrt{k}}\right)$
Complexity	$\max(\Theta(\frac{LR^2}{\epsilon}), \Theta(\frac{\sigma^2 R^2}{\epsilon^2}))$	$\max(\Theta(\sqrt{\frac{LR^2}{\epsilon}}), \Theta(\frac{\sigma^2 R^2}{\epsilon^2}))$,
Best Reach. Acc.	0	0

8.1.2 Objective function with weaker level of smoothness: Universal First-order methods (Chapters 4 and 5)

We have introduced and used intensively the notion of (δ, L)-oracle in order to represent inexactness in the first-order information of a smooth convex function. However, it appears that this notion can also be used in order to represent another difficulty as compared with the desired situation (a smooth convex function with exact first-order information). The exact oracle of a nonsmooth or a weakly smooth convex function can be seen as a particular case of (δ, L)-oracle.

As a consequence, the first-order methods initially developed for smooth convex problems (i.e. the PGM/DGM and FGM) can also be applied to nonsmooth and weakly smooth convex functions, obtaining in some sense universal first-order methods. Furthermore, the complexity of the PGM/DGM and FGM in these new situations can be directly derived from the behavior of these first-order methods when used with a (δ, L)-oracle.

The following table summarizes the complexities of the PGM/DGM when applied to smooth, weakly smooth or nonsmooth convex functions:

Objective function	Complexity PGM/DGM	Optimal ?
Smooth convex ($f \in F_L^{1,1}(Q)$)	$\Theta\left(\frac{LR^2}{\epsilon}\right)$	No
Weakly smooth convex ($f \in F_{L_\nu}^{1,\nu}(Q)$)	$\Theta\left(\left(\frac{L_\nu R^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+\nu}}\right)$	No
Nonsmooth convex ($f \in F_{L_0}^{0,0}(Q)$)	$\Theta\left(\frac{L_0^2 R^2}{\epsilon^2}\right)$	Yes

These methods are only optimal in the nonsmooth case, where the PGM is nothing else but the subgradient/mirror-descent method. On the other hand, the FGM is optimal for nonsmooth but also for weakly smooth and smooth convex problems:

Objective function	Complexity FGM	Optimal ?
Smooth convex ($f \in F_L^{1,1}(Q)$)	$\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$	Yes
Weakly smooth convex ($f \in F_{L_\nu}^{1,\nu}(Q)$)	$\Theta\left(\left(\frac{L_\nu R^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+3\nu}}\right)$	Yes
Nonsmooth convex ($f \in F_{L_0}^{0,0}(Q)$)	$\Theta\left(\frac{L_0^2 R^2}{\epsilon^2}\right)$	Yes

These results break the wall between smooth and nonsmooth convex optimization. Using the trick of the (δ, L)-oracle, we can obtain first-order methods

applicable to smooth, weakly smooth and nonsmooth convex functions. Furthermore, the FGM that has been criticized for its higher sensitivity with respect to oracle errors regains its superb when we are interested in convex problems with exact information but possibly with a weaker level of smoothness. The FGM when used with a well-chosen stepsize (depending on the level of smoothness of the objective function) can be seen as an universal optimal first-order method, reaching the optimal complexity in the smooth, weakly smooth and nonsmooth case.

Similar results can be obtained in the strongly convex case. Using here the notion of (δ, L, μ)-oracle, we have proved that the first-order methods initially designed for smooth strongly convex functions can also be applied to strongly convex functions with weaker level of smoothness (nonsmooth strongly convex function and weakly smooth strongly convex function) but also to uniformly convex functions with various levels of smoothness (smooth, nonsmooth or weakly smooth). In view of this result, we have derived the corresponding complexities of the PGM/DGM and FGM on these different classes.

8.1.3 Linearly Constrained problems : A Novel Double Smoothing Technique (Chapter 3)

The problem we are facing in Chapter 3 is not an inexactness of the first-order information, nor a lack of smoothness (at least for the initial formulation of the problem). The available first-order information is exact but, due to linear constraints, projections on the feasible set cannot be computed efficiently which makes the usual first-order methods non applicable.

A natural approach consists of dualizing the linear constraints, obtaining a dual problem which is unconstrained but non differentiable and applying to this dual problem a first-order method for nonsmooth convex problems like the subgradient method. However with such approach, we can only obtain an unattractive complexity $O\left(\frac{1}{\epsilon^2}\right)$ both for the primal and dual convergences.

Using the structure of the dual nonsmooth function, it is possible to apply the smoothing technique developed by Nesterov in [59] that transforms the nonsmooth convex dual function into a smooth convex approximation and applies to this function a fast gradient method. This approach allows us to solve the dual problem with a significantly better complexity proportional to $O(\frac{1}{\epsilon})$. However, the complexity improvement obtained with the classical smoothing technique is lost during the process of reconstructing a primal solution from the dual one. Because we cannot guarantee a fast decrease of the norm of the gradient, primal convergence (the one in which we are really interested) reduces to a complexity

not better than using the basic subgradient approach.

The double smoothing technique consists in going one step further in the smoothing process, making the dual objective not only smooth with a Lipschitz-continuous gradient but also strongly convex. Applying a fast gradient method for smooth strongly convex functions to this doubly smoothed dual objective function, we can solve the dual problem with accuracy ϵ in $O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations and from this nearly optimal dual solution, reconstruct a nearly optimal primal solution with the same level of accuracy.

Method	Dual function	Dual compl.	Primal compl.
Subgradient	Convex but Non-Smooth	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$
Simple Smoothing	Convex ∇ Lipschitz-cont	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$
Double Smoothing	Strongly convex ∇ Lipschitz-cont.	$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$

8.2 Directions for Further Research

After emphasizing the contribution of this thesis, let us conclude it by a discussion on different potential directions for further research:

8.2.1 First-order methods with inexact first-order information

Even if we have already gained an extensive understanding of the behavior of first-order of smooth convex optimization when used with inexact oracle, this thesis opens the door to some natural direction of further research that could extend the scope and applicability of our analysis:

- ◇ **Stochastic oracle in the strongly convex case**
  If we have considered separately strong convexity and stochasticity as two ways to improve the behavior of first-order methods with respect to errors, a natural extension is to combine them and analyze the effect of a (biased) stochastic oracle on smooth strongly convex functions.
- ◇ **Intermediate Gradient Method in the strongly convex case**
  Another natural direction for further research would be to extend the notion of Intermediate Gradient Method to the strongly convex case, trying

to develop a method with intermediate behavior (in term of the dependence in the condition number) between the slow but robust PGM/DGM and the fast but sensitive FGM. In particular, it would be nice to develop a very general Intermediate Gradient Method which is continuous in the strongly convex parameter (when $\mu = 0$, we retrieve the IGM already developed) and continuous with respect to the trade-off fastness/robustness (we recover the DGM and the FGM as extreme cases).

◇ **Variable and Approximable parameters for the inexact oracle**
Our analysis of the first-order methods with (δ, L)-oracle or with (δ, L, μ)-oracle assumes that the parameters δ,L and μ are constant over the successive iterations, or at least only takes into account the worst-case values for these quantities (i.e. the largest δ and L, the smallest μ). Furthermore, practical implementation of these methods requires for the moment an a priori knowledge of the parameters δ, L and μ.

It would be a real added value for the scope and applicability of these methods to adapt these first-order methods and their analyses in a way that allows to specify or estimate at each step the value of δ, L and μ for the current answer of the first-order oracle.

◇ **More optimistic analysis**
We considered in this thesis worst-case oriented results. Our behavior analysis of the first-order methods considers the worst-case situation of the worst objective function coupled with an inexact oracle that fight against the method, the successive errors combining with each other in the most damaging way. The average behavior with respect to errors of these method could be very different. We could expect that the errors will combine in a better way, canceling each other out, or a least in a neutral way, following a stochastic distribution without bias.

In particular, even if the accumulation of errors in the FGM is unavoidable in a worst-case analysis and can be seen on well-chosen numerical experiments, we could expect that there are many situations where this sensitivity to errors does not appear, where the FGM is robust and fast at the same time.

An on average analysis (theoretically or with extended numerical experiments) of the effect of inexact information on first-order methods is an interesting subject for further research.

◇ **Inexact oracle for second-order methods**
More generally, this thesis focus on first-order methods and it could be interesting to develop the same kind of results for second-order methods. A very promising direction for further research could be to adapt the definition of the (δ, L)-oracle by adding second-order information and

study the behavior of existing second-order methods, such as the Newton method or the Newton method with cubic regularization, when used with this inexact oracle. In particular, it would be very interesting to see if the faster method with cubic regularization is more sensitive to errors compared to the classical Newton method, i.e. whether there also exists a link between speed and sensitivity to errors for second-order methods.

8.2.2 Objective function with weaker level of smoothness: Universal First-order methods.

◇ **A self-adaptive universal optimal first-order method**
In this thesis, we have in some sense broken the wall between smooth convex optimization and nonsmooth convex optimization: first-order methods designed for smooth problems can be applied to objective functions with weaker levels of smoothness, provided that the stepsize is adapted in order to take into account the actual level of smoothness.

An interesting goal for further research is certainly to use these results in order to design a first-order method which is not only optimal for each level of smoothness separately but that can reach these optimal complexities without a priori knowledge of the level of smoothness. It would be very interesting to obtain a method able to adapt the Lipschitz-constant to the actual level of smoothness of the objective function and to exploit local differences of smoothness in some areas of the feasible set. If this online choice of the stepsize can be implemented, we could for example significantly accelerate the minimization of a nonsmooth function as long as we are far enough from the points where the function is non differentiable.

8.2.3 Linearly Constrained problems : A Novel Double Smoothing Technique.

◇ **Double Smoothing with inexact resolution of subproblems**
We have developed the double smoothing technique with the assumption that the subproblem defining the doubly smoothed dual objective function can be solved exactly, providing exact first-order information for the fast gradient method applied to this function. This assumption means that the only difficulty in the initial problem comes from the linear constraint (i.e. without this constraint the problem is in some sense trivial). Using our extended analysis of the effect of inexact first-order information on the fast-gradient method, it could be possible for further research to drop this assumption, linking in some sense the two parts of this thesis.

Indeed, the doubly smoothed dual objective function $\theta_{\rho,\mu,\kappa}(.)$ is a max type function of the form studied in subsection 4.2.2. We know that solving the subproblems defining $\theta_{\rho,\mu,\kappa}(.)$, up to an accuracy of order δ, provides a (δ, L, κ)-oracle for this function. Furthermore, in the double smoothing technique, we apply to this objective function a FGM for smooth strongly convex functions, a method for which the behavior, when used with a (δ, L, κ)-oracle, has been studied in section 5.5. Putting together these results, we can determine to which level of accuracy we need to solve the different subproblems in order to obtain a final target accuracy of order ϵ for the doubly smoothed dual objective function $\theta_{\rho,\mu,\kappa}(.)$.

However, we have seen in Chapter 3 that controlling the gap $\theta_{\rho,\mu,\kappa}(y_k) - \theta^*_{\rho,\mu,\kappa}$ is not sufficient for ensuring an accuracy of order ϵ for the primal objective function. All the necessary materials and tools have been developed in this thesis but some technical work must still be done for this analysis in order to obtain a complete complexity analysis of the double smoothing technique without assuming exact resolution of subproblems.

Similarly, the results obtained in this thesis open a possibility to develop complete complexity analysis, taking into account outer and inner complexities, for various dual-type approaches such as Augmented Lagrangians.

- ◇ **Inexact Projection: general analysis of the effects on first-order methods**
  If this thesis considers the situation where projections on the feasible set are difficult due to linear constraints and tackle this difficulty by dualizing these constraints, it could be also interesting to consider the general effect of inexact projections on first-order methods. Introducing a notion of inexact projection and coupling it with the (δ, L)-oracle, we could then obtain a very general framework for analyzing the behavior of the first-order methods in the presence of inexactness at all levels of the algorithm (inexact first-order information and inexact projection on the feasible set).

Table of Methods

Primal Gradient Method (PGM)

◇ Exact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	Exist: [58]	Exist: [58]	Exist [62]
Non Euclidean Setup:	New: Sections 2.4.2, 7.2	Further Research	New: Section 7.2

◇ Inexact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	New: Section 4.3.1	New: Section 5.3	New: 7.2
Non Euclidean Setup:	New: Section 7.2	Further Research	New: Section 7.2

◇ Inexact Stochastic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	Exist: [44] + New Stepsizes and Biased Case: Section 7.2	Exist: [28]	Exist : [67] + New Stepsizes and Biased Case: Section 7.2
Non Euclidean Setup:	Exist: [44] + New Stepsizes and Biased Case: Section 7.2	Further Research	New: Section 7.2

Dual Gradient Method (DGM)

◇ Exact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	Exist: [62]	New: Section 5.4	Exist: [62]
Non Euclidean Setup:	New: Sections 2.4.4, 7.4	Further Research	New: Section 7.4

◇ Inexact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	New: Section 4.3.2	New: Section 5.4	New: Section 7.4
Non Euclidean Setup:	New: Section 7.4	Further Research	New: Section 7.4

◇ Inexact Stochastic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	New: Section 7.4	Further Research	New: Section 7.4
Non Euclidean Setup:	New: Section 7.4	Further Research	New: Section 7.4

Fast Gradient Method (FGM)

◇ Exact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	Exist: [58, 59]	Exist: [58] + New version: Section 5.5	Exist [62]
Non Euclidean Setup:	Exist [59]	Further Research	New: Section 7.5

◇ Inexact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	New: Section 4.4	New: Section 5.5	New: Section 7.5
Non Euclidean Setup:	New: Section 7.5	Further Research	New: Section 7.5

◇ Inexact Stochastic Oracle

Objective Function.	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	Exist: [44] + New Version and Biased Case: Section 7.5	Exist: [28]	Exist : [67] + New Version and Biased Case: Section 7.5
Non Euclidean Setup:	Exist: [44] + New Version and Biased Case: Section 7.5	Further Research	New: Section 7.5

Intermediate Gradient Method (IGM)

◇ Inexact Deterministic Oracle

Objective Function:	Smooth Convex	Smooth Strongly Conv.	Composite: Smooth Convex + Easy Nonsmooth
Euclidean Setup:	New: Section 6.1.1	Further Research	Further Research
Non Euclidean Setup:	New: Section 6.1.2	Further Research	Further Research

Bibliography

[1] P.J. Antsaklis and A.N. Michel. *Linear Systems*. Birkhauser Book (2006)

[2] M. Baes. Estimate sequence methods: extensions and approximations. *IFOR Internal report, ETH Zurich, Switzerland* (2009)

[3] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* , **31(3)** , 167-175 (2009).

[4] A. Beck and M. Teboulle. A Fast Iterative Shrinkage Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal of Imaging Sciences*, **2(1)**, 183-202 (2009).

[5] A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery. In Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar Eds., pp. 33–88. Cambribge University Press, 2010.

[6] S. Becker, E. Candès and M. Grant. Templates for Convex Cone Problems with Applications to Sparse Signal Recovery. *Mathematical Programming Computation* **3(3)**, 165–218 (2010).

[7] S. Becker, J. Bobin and E.J. Candès. NESTA: a fast and accurate first-order method for sparse recovery, *SIAM J. Imaging Sciences* **4(1)**, 1-39 (2009).

[8] P. Bernhards and A. Rapaport. On a theorem of Danskin with an application to a theorem of Von Neumann-Sion. *Nonlinear analysis*, **24**, 1163-1181 (1995).

[9] A. Ben-Tal and A. Nemirovsk. *Lectures on Modern Convex Optimization: Anal- ysis, Algorithms, Engineering Applications.*. MPS-SIAM Series on Optimization, Philadelphia: SIAM (2000)

[10] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific. (2009)

[11] J.F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer. (2000)

[12] J.M. Borwein and A.S. Lewis. Duality Relationship for Entropy-Like Minimization Problems. *SIAM. J. Control and Optimization*, **29(2)**, 325-338 (1991).

[13] J.M. Borwein and A.S. Lewis. Partially finite convex programming, Part I: Quasi relative interiors and duality theory. *Mathematical Programming*, **57**, 15-48 (1992).

[14] J.M. Borwein and A.S. Lewis. Partially finite convex programming, Part II: Explicit Lattice models. *Mathematical Programming*, **57**, 49-83 (1992).

[15] J.M. Borwein and A.S. Lewis Partially finite programming in L_1 and the existence of maximum entopy estimates. *SIAM. J. Optimization*, **3(2)**, 248-267 (1993).

[16] J.M. Borwein and H. Wolkowicz. A simple constraint qualification in infinite dimensional programming. *Mathematical Programming*, **35**, 83-96 (1986).

[17] R.I. Bot, S-M. Grad and G. Warka New regularity conditions for Lagrange and Fenchel-Lagrange duality in infinite dimensional spaces. *Mathematical Inequalities and Applications*,**12(1)**, 171-189 (2009).

[18] S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, SIAM, (1994)

[19] S. Boyd and L. Vandenbergh. *Convex Optimization (7th printing with corrections)*. Cambdrige University Press. (2009)

[20] R. Correa and C. Lemarechal. Convergence of some algorithms for convex minimization. *Mathematical Programming, Serie A*,**62**, 261-275 (1993).

[21] A. D'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal of Optimization*, **19(3)**, 1171-1183 (2008).

[22] J.M. Danskin. *The theory of Max-Min and its application to weapons allocation problems*. Springer-Verlag (1967).

[23] O. Devolder, F. Glineur and Y. Nesterov. Double Smoothing Technique for Large-Scale Linearly Constrained Convex Optimization. *SIAM Journal of Optimization*, **22(2)** , (2012)

[24] O. Devolder, F. Glineur and Yu. Nesterov. First-order Methods of Smooth Convex Optimization with Inexact Oracle. *Mathematical Programming, Serie A, Accepted*, (2013).

[25] O. Devolder. Stochastic First Order Methods in Smooth Convex Optimization. *CORE Discussion Paper 2011/70*, (2011).

[26] O. Devolder, F. Glineur and Yu. Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Paper 2013/16*, (2013).

[27] O. Devolder, F. Glineur and Yu. Nesterov. Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle. *CORE Discussion Paper 2013/17*, (2013).

[28] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. *SIAM Journal on Optimization, Accepted*, (2012)

[29] M. Hintermuller. A proximal bundle method based on approximative subgradient. *Computational Optimization and Applications*, **20**, 245-266 (2001)

[30] C. Hu, J.T. Kwok and W. Pan.. Accelerated Gradient Methods for Stochastic Optimization and Online Learning. *Neural Information Processing Systems (NIPS), Vancouver, Canada*, (2009).

[31] V. Jeyakumar, N. Dinh and G.M. Lee A New closed cone constraint qualification for convex optimization. *Applied Mathematics Report AMR 04/08, University of New South Wales, Australia* (2004).

[32] V. Jeyakumar and H. Wolkowicz. Generalizations of Slater's constraint qualification for infinite convex programs. *Mathematical Programming*, **57**, 85-101 (1992).

[33] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *Technical Report*, (2010).

[34] A. Juditsky, K. Karzan and A. Nemirovski. l_1 minimization via randomized first order algorithms. *submitted to Mathematical programming, Serie A*, (2010).

[35] A. Juditsky and A. Nemirovski. Large Deviations of Vector-Valued Martingales in 2-smooth Normed Spaces. *Submitted to the Annals of Probability.*, (2008).

[36] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *In: S. Sra, S. Nowozin, S. Wright, Eds., Optimization for Machine Learning*, The MIT Press, (2011).

[37] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: Utilizing Problem's Structure *In: S. Sra, S. Nowozin, S. Wright, Eds., Optimization for Machine Learning*, The MIT Press, (2011).

[38] L.Khachiyan, S.Tarasov, and E.Erlich. The inscribed ellipsoid method. *Soviet Math. Dokl. (In Russian)* , **298** (1988).

[39] L. Kachiyan, A Nemirovski and Y. Nesterov. Optimal methods of convex programming and polynomial methods of linear programming. *In H. Elster, editor, Modern Mathematical Methods of Optimization*, Akademie Verlag 75-115 (1993).

[40] D.E. Kirk. *Optimal Control: An Introduction*. Dover publication, Inc. (2004).

[41] K. Kiwiel. A proximal bundel method with approximative subgradient linearization. *SIAM Journal of Optimization*, **16(4)**, 1007-1023 (2006)

[42] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems*. Wiley Interscience (1972)

[43] G. Lan, Z. Lu and R. D.C Monteiro. Primal-dual first-order methods with $O(\frac{1}{\epsilon})$ iteration-complexity for cone programming. *Mathematical Programming*, **126(1)**, 1-29, (2011).

[44] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming Serie A*, **133(1-2)**, 365-397, (2012)

[45] G. Lan, A. Nemirovski and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming Serie A*, **134(2)**, 425-458, (2012)

[46] C. Lemaréchal. Nonsmooth optimization and descent methods. *Research Report 78-4, IIASA, Laxenburg, Austria* (1978).

[47] C. Lemaréchal, J. J. Strodiot, and A. Bihain. On a bundle algorithm for nonsmooth optimization. *in O.L. Mangasarian, R.R.Meyer, and S.M. Robinson, Eds. Nonlinear Programming 4, Academic Press, pages 245-282* (1981).

[48] C. Lemaréchal, A. Nemirovski, and Yu. Nesterov. New variants of bundle methods. *Mathematical Programming Ser. B* **69(1-3)**, 111-147 (1995).

[49] I. Necoara and J.A.K. Suykens. Application of a Smoothing Technique to Decomposition in Convex Optimization. *IEEE Transactions on Automatic Control*, **53**, 2674-2679 (2008).

[50] A. Nedic and D.Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical programming, Serie A*, **125(1)**, 75-99 (2010).

[51] A.Nemirovski. Prox-method with rate of convergence $o(\frac{1}{t})$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *Siam Journal of Optimization*, **15(1)**, 229-251 (2004).

[52] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *Siam Journal of Optimization*, **19(4)**, 1574-1609 (2009).

[53] A. Nemirovski and Y.Nesterov. Optimal methods for smooth convex minimization. *Zh. Vichisl. Mat. Fiz. (In Russian)*, **25(3)**, 356-369 (1985).

[54] A. Nemirovski and Yu.Nesterov. *Interior point polynomial methods in convex program- ming: Theory and Applications*. SIAM, Philadelphia (1994).

[55] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley (1983)

[56] Yu. Nesterov. A method for unconstrained convex minimization with the rate of convergence of $O(\frac{1}{k^2})$, *Doklady AN SSSR*, **269**, 543-547 (1983).

[57] Yu. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex function, *Èkonom. i. Mat. Metody (In Russian)*, **24**, 509-517 (1988).

[58] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers (2004)

[59] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Serie A*, **103(1)**, 127-152 (2005).

[60] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *Siam Journal of Optimization*, **16(1)**, 235-249 (2005).

[61] Yu. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming A*, **110(2)**, 245-259 (2007).

[62] Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, **76**, (2007)

[63] Yu. Nesterov. Primal-Dual subgradient methods for convex problems. *Mathematical programming, Serie B*, **120(1)**, 221-259 (2009).

[64] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, **22(2)**, 341-362 (2012)

[65] Yu. Nesterov. Subgradient methods for huge-scale optimization problems. *CORE Discussion Paper*, **1**, (2012)

[66] J. Nocedal and S. Wright. *Numerical Optimization*. 2nd edition, Springer (2006)

[67] Q. Lin, X. Chen and J. Pena. A Smoothing Stochastic Gradient Method for Composite Optimization. *Manuscript: arXiv:1008.5204v2*, (2010).

[68] B.T. Polyak. *Introduction to Optimization*. Optimization Software Inc, (1987)

[69] T. Pennanen. Introduction to convex optimization in financial markets. *Mathematical Programming Serie B*, **134(1)**, 157-186, (2012)

[70] D. P. Palomar and Y. C. Eldar, Eds., *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, (2009).

[71] P. Richtarik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Accepted in Mathematical Programming*, (2012)

[72] K. Scheinberg, D. Goldfarb and X. Bai Fast first-order methods for composite convex optimization with line search. Optimization Online, (2011)

[73] A. Shapiro, D. Dentcheva and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM Series on Optimization, (2009)

[74] N.Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer Series in Computational Mathematics. Springer-Verlag (1985).

[75] M. Schmidt, N. Le Roux and F. Bach. Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization. INRIA, Preprint, (2011).

[76] S. Sra, S. Nowozin and S. J. Wright, Eds. *Optimization for Machine Learning*, The MIT Press, (2011)

[77] P. Tseng. On accelerated Proximal Gradient Methods for Convex-Concave Optimization *Submitted to SIAM Journal on Optimization* (2008).

[78] E. Zeidler. *Nonlinear Functional Analysis and its Application III: Variational Methods and Optimization*, Springer-Verlag (1985).

[79] T. Zhou, D. Tao and X. Wu. NESVM: A Fast Gradient Method for Support Vector Machines, *IEEE International Conference on Data Mining*, pp.679-688 (2010).