CORE DISCUSSION PAPER

2013/17

Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle

O. Devolder, F. Glineur and Yu. Nesterov. *

April 22, 2013

Abstract

Between the robust but slow (primal or dual) gradient methods and the fast but sensitive to errors fast gradient methods, our goal in this paper is to develop first-order methods for smooth convex problems with intermediate speed and intermediate sensitivity to errors.

We develop a general family of first-order methods, the Intermediate Gradient Method (IGM), based on two sequences of coefficients. We prove that the behavior of such kind of method is directly governed by the choice of coefficients and that the existing dual and fast gradient methods can be retrieved with particular choices for the coefficients. Moreover, the degree of freedom in the choice of these coefficients can be also used in order to generate intermediate behaviors.

We propose a switching policy for the coefficients that allows us to see the corresponding IGM as a smart switching between fast and dual gradient methods and to reach target accuracies, unreachable by the fast gradient methods, in a significantly smaller number of iterations compared to what is needed using the slow gradient methods. With another choice for the coefficients, we are also able to generate methods exhibiting the full spectrum of convergence rates, corresponding to every possible trade off between fastness of the method and robustness to errors.

*Center for Operations Research and Econometrics (CORE), Université catholique de Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; E-mail: Olivier.Devolder@uclouvain.be.
This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The first author is a F.R.S.-FNRS Research Fellow. The research of the third author was partly supported by the grant 'Action de recherche concertée ARC 04/09-315' from the Direction de la recherche scientifique - Communautée francaise de Belgique.' The third author also acknowledges the support from Laboratory of Structural Methods of Data Analysis in Predictive Modelling, through the RF government grant 11.G34.31.0073. The scientific responsibility rests with its authors.

1 Introduction

We consider the following convex optimization problem:

$$f^* = \min_{x \in Q} f(x), \tag{1.1}$$

where Q is a closed convex set in a finite-dimensional space E, and function f is convex on Q. We assume that problem (1.1) is solvable with optimal solution x^*.

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.

1.1 Exact and Inexact Oracle

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 (see for example [7]) that functions belonging to this class satisfy

$$0 \;\leq\; f(x) - \big(f(y) + \langle \nabla f(y), x - y\rangle\big) \;\leq\; \frac{L}{2}\|x - y\|_E^2 \text{ for all } x, y \in Q. \tag{1.2}$$

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

$$0 \;\leq\; f(x) - \big(f_L(y) + \langle g_L(y), x - y\rangle\big) \;\leq\; \frac{L}{2}\|x - y\|_E^2 \text{ for all } x \in Q.$$

Our definition of an inexact oracle, introduced in [2], simply consists in introducing a given amount δ of tolerance in this pair of inequalities:

Definition 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

$$0 \;\leq\; f(x) - \big(f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y\rangle\big) \leq\; \frac{L}{2}\|x - y\|_E^2 + \delta \text{ for all } x \in Q. \tag{1.3}$$

Constant δ is called the *accuracy* of the oracle. The oracle is exact when $\delta = 0$ and inexact when $\delta > 0$.

We have shown in [2] that this definition can be used to represent various natural situations where only inexact first-order information is available. Let us recall here the more important examples:

- **Computation 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. Then $f_{\delta,L}(y) \stackrel{\text{def}}{=} f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y}\rangle$, and $g_{\delta,L}(y) \stackrel{\text{def}}{=} \nabla f(\hat{y})$ is a (δ, L)-oracle with $\delta = M\|y - \hat{y}\|_E^2, \quad L = 2M$.
- **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 $\big\|\nabla f(y) - \tilde{\nabla} f_y\big\|_E^* \leq \Delta_2$.
  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 $L = M$.

- **Inexact resolution of subproblems for max-type functions**
  Let us consider smooth convex optimization problems of the form (1.1) whose objective function $f \in F_M^{1,1}(Q)$ is defined by another optimization problem:

$$f(x) \;=\; \max_{u \in U} \Psi(x,u), \tag{1.4}$$

  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. In [2], we have considered three classes of max-type functions for which approximate solution of subproblem (1.4) allows the construction of a (δ, L)-oracle:

  1. *Functions obtained by smoothing techniques*
     Let
$$\Psi(x,u) = G(u) + \langle Au, x\rangle,$$
     where $A : F \to E^*$ is a linear operator, and $G(u)$ is a differentiable, strongly concave function with concavity parameter $\kappa > 0$. The importance of this class of functions is justified by the smoothing approach for non-smooth convex optimization (see [8, 9, 10, 1]).
     Suppose that for all $y \in Q$ we can find a point $u_y \in U$ satisfying condition
$$\Psi(y, u_y^*) - \Psi(y, u_y) \;\leq\; \tfrac{\delta}{2}. \tag{1.5}$$
     then the pair $(\Psi(y,u_y), Au_y)$ corresponds to an (δ, L)-oracle with $L = \frac{2}{\kappa}\|A\|^2_{F\to E^*}$ (where $\|A\|_{F\to E^*} = \max\{\|Au\|_{E^*} : \|u\|_F = 1\}$).
  2. *Moreau-Yosida Regularization*
     Let us consider functions of the form
$$f(x) \;=\; \min_{u\in U}\left\{\mathcal{L}(x,u) \stackrel{\text{def}}{=} h(u) + \tfrac{\kappa}{2}\|u-x\|_2^2\right\}, \tag{1.6}$$
     where h is a smooth convex function on a convex set $U \subset \mathbb{R}^n$ endowed with the usual Euclidean norm $\|u\|_2^2 = \langle u, u\rangle$.
     Instead of solving exactly problem (1.6), we compute a feasible solution u_x satisfying
$$\max_{u\in U}\left\{\mathcal{L}(x,u_x) - \mathcal{L}(x,u) + \frac{\kappa}{2}\|u-u_x\|_2^2\right\} \;\leq\; \delta.$$
     Then for all $x \in Q$ the objects
$$\begin{aligned} f_{\delta,L}(x) \;&=\; \mathcal{L}(x,u_x) - \delta \;=\; h(u_x) + \tfrac{\kappa}{2}\|u_x - x\|_2^2 - \delta, \\ g_{\delta,L}(x) \;&=\; \nabla_1\mathcal{L}(x,u_x) \;=\; \kappa(x-u_x) \end{aligned} \tag{1.7}$$
     correspond to an answer of an (δ, L)-oracle with $L = \kappa$.
  3. *Functions defined by Augmented Lagrangians*
     Consider the following convex problem: $\max\limits_{u\in U}\{h(u) : Au = 0\}$ 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 $\min\limits_{x\in E} f(x)$ where
$$f(x) \;\stackrel{\text{def}}{=}\; \max_{u\in U}\left[\Psi(x,u) \stackrel{\text{def}}{=} h(u) + \langle Au, x\rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\right]. \tag{1.8}$$

Assume that, instead of solving (1.8) exactly, we compute an approximate solution $u_x \in U$ such that

$$\max_{u \in U} \langle \nabla_2 \Psi(x, u_x), u - u_x \rangle \leq \delta.$$

Then the objects

$$f_{\delta,L}(x) = \Psi(x, u_x), \quad g_{\delta,L}(x) = \nabla_1 \Psi(x, u_x) = A u_x \tag{1.9}$$

correspond to a (δ, L)-oracle with $L = \frac{1}{\kappa}$.

1.2 First-order methods using exact and inexact oracles

In smooth convex optimization, first-order methods can be divided in two different families.

On one hand, we have the Primal Gradient Method (PGM) (see [7]) and the Dual Gradient Method (DGM) (see [11]). Applied to a function $f \in F_L^{1,1}(f)$ endowed with an exact oracle (i.e. $\delta = 0$), these methods exhibits a convergence rate of the form:

$$f(y_k) - f^* \leq O(1) \frac{LR^2}{k}$$

where R is the distance between the initial iterate and the optimal solution set of problem (1.1). It means that reaching a target accuracy ϵ takes $\Theta\left(\frac{LR^2}{\epsilon}\right)$ iterations.

Remark 1 As the primal and dual gradient methods shares their main theoretical properties, we often use the generic denomination Gradient Method (GM).

On the other hand, we have a family of accelerated gradient methods, the Fast Gradient Methods (FGM), developed by Nesterov in different variants since 1983 (see [5, 6, 7, 8]). These methods applied to a smooth convex function with exact oracle exhibit a much better convergence rate of the form:

$$f(y_k) - f^* \leq O(1) \frac{LR^2}{k^2}.$$

An ϵ-solution can be obtained after only $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations. The complexity $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ (and the corresponding convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$) is in fact optimal for a first-order method applied to a function $f \in F_L^{1,1}(Q)$ (see [4, 7]). It is impossible to obtain a method with a better behavior than the FGM in the exact case.

Remark 2 Although all the fast gradient type methods exhibits the same order of complexity, in this paper, we denotes, unless stated otherwise, by FGM the latest variant developed by Nesterov in [8].

In view of this difference of complexities, it is clear that in the exact case (i.e. when the oracle for f is exact), the FGM outperforms clearly the GM and can be seen as the first-order method of choice.

In the inexact case, when the function f is endowed with a (δ, L) oracle with accuracy δ, the situation is more complicated.

- 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 becomes (see [2]):

$$f(y_k) - f^* \leq O(1)\frac{LR^2}{k} + \delta.$$

  The method is slow but there is no accumulation of error. 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 proportional to $\frac{LR^2}{\epsilon-\delta}$.

- The FGM can be seen as a fast method but sensitive with respect to oracle error. Indeed, when used with a (δ, L) oracle, the FGM exhibits the convergence rate (see [2]):

$$f(y_k) - f^* \leq O(1)\frac{LR^2}{k^2} + O(1)k\delta.$$

  Contrarily to the GM, the use of inexact oracle in FGM results in errors 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 $\Theta\left(\sqrt[3]{\frac{LR^2}{\epsilon}}\right)$ 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 not so 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}.

1.3 What can we expect ?

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. The following theorem obtained in [2], show us that the accumulation of errors is an intrinsic and unavoidable property of any fast first-order method using inexact oracle:

Theorem 1 Consider a first-order method for $F^{1,1}_L(Q)$ with convergence rate $O(\frac{LR^2}{k^p})$ in the exact case ($1 \leq p \leq 2$). Assume that the bounds on the performance of this method, as applied to a problem equipped with inexact (δ, L)-oracle, are given by inequality

$$f(z_k) - f^* \quad \leq \quad O(1)\frac{LR^2}{k^p} + O(1)k^q\delta.$$

Then $q \geq p - 1$.

In particular, this theorem show us that:

- $q = 0 \Rightarrow p \leq 1$: The GM is slow but it is the fastest first-order method that avoid error accumulation.
- $p = 2 \Rightarrow q \geq 1$: Any first-order method with optimal convergence rate $\Theta\left(\frac{LR^2}{k^2}\right)$ must suffer from error accumulation and FGM has the lowest possible rate of error accumulation for such a method: $\Theta(k\delta)$.

This result is not a 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: faster the method is, higher the sensitivity to error is.

But, 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 paper, we want to develop methods with intermediate behavior between GM and FGM methods.

1.4 Paper structure

The structure of this paper 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 3, we see that the existing DGM and FGM are nothing else that an IGM but with particular choices of the coefficients. In Section 4, 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 satisfies and derive lower bound on the complexity that we can expect with an optimal choice of the coefficients.

In Section 5 and 6, 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 7, with another choice for the coefficients, we are able to generate methods exhibiting the whole spectrum of convergence rates given by Theorem 1, corresponding to every possible trade-off between fastness of the method and robustness to errors. In the last section (Section 8), we present a numerical illustration of the results obtained in this paper.

2 The Intermediate Gradient Method (IGM)

2.1 General Intermediate Gradient Method

Let $\{\alpha_i\}_{i\geq 0}$ and $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying for all $i \geq 0$

$$\alpha_i^2 \leq B_i \leq \sum_{j=0}^{i} \alpha_j \tag{2.1}$$

$$0 \leq \alpha_i \leq B_i. \tag{2.2}$$

We define also $A_i = \sum_{j=0}^{i} \alpha_j$ and $\tau_i = \frac{\alpha_{i+1}}{B_{i+1}}$.

Let $d(x)$ be a prox-function i.e. a differentiable and strongly convex function on Q, and let $x_0 = \arg\min_{x \in Q} d(x)$ be its prox-center.

Translating and scaling d if necessary, we can always ensure that

$$d(x_0) = 0, \quad d(x) \geq \frac{1}{2} \|x - x_0\|_E^2, \quad \forall x \in Q. \tag{2.3}$$

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 1 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

$$w_k = \arg\min_{x \in Q} \{\frac{L}{2} \|x - x_k\|_E^2 + \langle g_{\delta,L}(x_k), x - x_k \rangle\} \tag{2.4}$$

5: Compute

$$z_k = \arg\min_{x \in Q} \{L d(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle\} \tag{2.5}$$

6: Compute

$$y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k \tag{2.6}$$

7: Compute $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
8: **end for**

Remark 3 We have $B_0 = A_0 = \alpha_0$ and therefore $y_0 = w_0$. We do not need to define y_{-1}.

Remark 4 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 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

$$\begin{aligned}
\Psi_0^* &= \min_{x\in Q}\{Ld(x) + \alpha_0[f_0 + \langle g_0, x - x_0\rangle]\} \\
&\overset{(2.3)}{\geq} \alpha_0 \min_{x\in Q}\{f_0 + \langle g_0, x - x_0\rangle + \frac{L}{2}\|x - x_0\|_E^2\} \\
&\overset{(1.3)}{\geq} \alpha_0 f(y_0) - \alpha_0\delta.
\end{aligned}$$

  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

$$\Psi_{k+1}^* \overset{(2.5)}{\geq} \Psi_k^* + \min_{x\in Q}\{\frac{L}{2}\|x - z_k\|_E^2 + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle]\}. \tag{2.7}$$

  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

$$\begin{aligned}
& \Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
\geq\ & A_k f(y_k) - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
=\ & (A_{k+1} - B_{k+1})f(y_k) - E_k + (B_{k+1} - \alpha_{k+1})f(y_k) \\
& + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
\overset{(1.3)}{\geq}\ & (A_{k+1} - B_{k+1})f(y_k) - E_k + (B_{k+1} \\
& - \alpha_{k+1})[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] \\
& + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\
=\ & (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} + \\
& \langle g_{k+1}, (B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle.
\end{aligned}$$

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

$$\begin{aligned}\Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ \geq (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1} f_{k+1} + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle. \quad (2.8)\end{aligned}$$

Therefore using the equations (2.7) and (2.8), 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

$$\begin{aligned}\Psi_{k+1}^* &\geq (A_{k+1} - B_{k+1}) f(y_k) - E_k \\ &\quad + B_{k+1} \min_{y \in Q}\{f_{k+1} + \langle g_{k+1}, y - x_{k+1}\rangle + \frac{L}{2}\|y - x_{k+1}\|_E^2\} \\ &\overset{(1.3),(2.4)}{\geq} (A_{k+1} - B_{k+1}) f(y_k) - E_k + B_{k+1}(f(w_{k+1}) - \delta) \\ &\overset{(2.6)}{\geq} A_{k+1} f(y_{k+1}) - E_k - B_{k+1}\delta = A_{k+1} f(y_{k+1}) - E_{k+1}.\end{aligned}$$

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 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 2.1 and 2.2. 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 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.

2.2 Variant with prox-type subproblems

The intermediate gradient method presented in the subsection 2.1 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 y_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 defined by

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

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

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

We propose the following modification of the IGM:

Algorithm 2 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

$$y_0 = \arg\min_{x \in Q} \{Ld(x) + \alpha_0 \langle g_{\delta,L}(x_0), x - x_0 \rangle\} \tag{2.11}$$

4: **for** k = 0 : ... **do**
5: Compute

$$z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle\} \tag{2.12}$$

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

$$\hat{x}_{k+1} = \arg\min_{x \in Q} \{LV(x, z_k) + \alpha_{k+1} \langle g_{\delta,L}(x_{k+1}), x - z_k \rangle\} \tag{2.13}$$

9: Compute $w_{k+1} = \tau_k \hat{x}_{k+1} + (1 - \tau_k) y_k$.
10: Compute

$$y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}. \tag{2.14}$$

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.

Remark 5 As an example, let us consider the situation where $Q = \{x \in \mathbb{R}^n_+ | \sum_{i=1}^n x^{(i)} = 1\}$, $\|.\|_E = \|.\|_1 = \sum_{i=1}^n \left|x^{(i)}\right|$ and $d(x) = \ln(n) + \sum_{i=1}^n x^{(i)} \ln(x^{(i)})$ (entropy distance). In this case, 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 [8]). 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 Section 5.3 in [8]):

$$x^i_{opt} = \frac{z^i_k \exp(-\frac{\alpha_{k+1}}{L} \nabla f(x_{k+1})^i)}{\sum_{j=1}^n z^j_k \exp(-\frac{\alpha_{k+1}}{L} \nabla f(x_{k+1})^j)}, \quad i = 1, ..., n. \tag{2.15}$$

Furthermore, the convergence rate of this modified IGM is the same than for the basic IGM:

Lemma 2 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

$$\begin{aligned}\Psi_0^* &\overset{(2.11)}{=} Ld(y_0) + \alpha_0[f_0\langle g_0, y_0 - x_0\rangle] \\ &\overset{(2.3)}{\geq} \alpha_0[f_0 + \langle g_0, y_0 - x_0\rangle + \frac{L}{2}\|y_0 - x_0\|_E^2] \\ &= \alpha_0 f(y_0) - \delta B_0.\end{aligned}$$

- 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

$$\begin{aligned}\Psi_{k+1}(x) &= Ld(x) + \sum_{i=0}^{k+1} \alpha_i[f_i + \langle g_i, x - x_i\rangle] \\ &\overset{(2.9)}{=} LV(x, z_k) + Ld(z_k) + \langle L\nabla d(z_k), x - z_k\rangle \\ &\quad + \sum_{i=0}^{k+1} \alpha_i[f_i + \langle g_i, x - x_i\rangle] \\ &\geq LV(x, z_k) + Ld(z_k) + \sum_{i=0}^{k} \alpha_i[f_i + \langle g_i, z_k - x_i\rangle] \\ &\quad + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ &\overset{(2.12)}{=} \Psi_k^* + LV(x, z_k) + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle].\end{aligned}$$

  On the other hand, we have, assuming $\Psi_k^* \geq A_k f(y_k) - E_k$:

$$\begin{aligned}&\Psi_k^* + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ \geq\ & A_k f(y_k) - E_k + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ =\ & (A_{k+1} - B_{k+1})f(y_k) - E_k + (B_{k+1} - \alpha_{k+1})f(y_k) \\ & + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ \overset{(1.3)}{\geq}\ & (A_{k+1} - B_{k+1})f(y_k) - E_k \\ & + (B_{k+1} - \alpha_{k+1})[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle] \\ & + \alpha_{k+1}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle] \\ =\ & (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} \\ & + \langle g_{k+1}, (B_{k+1} - \alpha_{k+1})(y_k - x_{k+1}) + \alpha_{k+1}(x - x_{k+1})\rangle.\end{aligned}$$

  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

$$\begin{aligned}
\Psi^*_{k+1} &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} \\
&\quad + \min_{x \in Q}\{LV(x, z_k) + \alpha_{k+1}\langle g_{k+1}, x - z_k\rangle\} \\
&\overset{(2.13)}{=} (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}f_{k+1} \\
&\quad LV(\hat{x}_{k+1}, z_k) + \alpha_{k+1}\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle \\
&= (A_{k+1} - B_{k+1})f(y_k) - E_k \\
&\quad + B_{k+1}[f_{k+1} + \frac{L}{B_{k+1}}V(\hat{x}_{k+1}, z_k) + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle] \\
&\overset{(2.10)}{\geq} (A_{k+1} - B_{k+1})f(y_k) - E_k \\
&\quad + B_{k+1}[f_{k+1} + \frac{L}{2B_{k+1}}\|\hat{x}_{k+1} - z_k\|^2_E + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle].
\end{aligned}$$

As $\alpha^2_{k+1} \leq B_{k+1}$, we have $\frac{1}{B_{k+1}} \geq \tau^2_k$ and therefore

$$\begin{aligned}
&\Psi^*_{k+1} \geq (A_{k+1} - B_{k+1})f(y_k) - E_k \\
&+ B_{k+1}[f_{k+1} + \tau_k\langle g_{k+1}, \hat{x}_{k+1} - z_k\rangle + \frac{\tau^2_k L}{2}\|\hat{x}_{k+1} - z_k\|^2_E].
\end{aligned}$$

But

$$\tau_k(\hat{x}_{k+1} - z_k) = w_{k+1} - x_{k+1}$$

and we obtain

$$\begin{aligned}
\Psi^*_{k+1} &\geq (A_{k+1} - B_{k+1})f(y_k) - E_k \\
&\quad + B_{k+1}[f_{k+1} + \langle g_{k+1}, w_{k+1} - x_{k+1}\rangle + \frac{L}{2}\|w_{k+1} - x_{k+1}\|^2_E] \\
&\overset{(1.3)}{\geq} (A_{k+1} - B_{k+1})f(y_k) - E_k + B_{k+1}[f(w_{k+1}) - \delta] \\
&\overset{(2.14)}{\geq} A_{k+1}f(y_{k+1}) - E_k - B_{k+1}\delta.
\end{aligned}$$

□

We can now apply the Theorem 2 to this modified IGM and conclude that it exhibits exactly the same convergence rate as the original IGM developed in subsection 2.1.

3 Link with existing methods

3.1 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 [8]:

Algorithm 3 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

$$y_k = \arg\min_{x \in Q} \{ \frac{L}{2} \|x - x_k\|_E^2 + \langle g_{\delta,L}(x_k), x - x_k \rangle \} \tag{3.1}$$

5: Compute

$$z_k = \arg\min_{x \in Q} \{ L d(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle \} \tag{3.2}$$

6: Compute

$$x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k. \tag{3.3}$$

7: **end for**

This method exhibits the following convergence rate (see Theorem 2 with $B_i = A_i$ for all $i \geq 0$)

$$f(y_k) - f^* \leq \frac{L d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} A_i}{A_k} \delta. \tag{3.4}$$

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 [8], 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 [2].

Compared to this existing scheme developed in [8], 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 1) due to the condition $\alpha_k^2 \leq B_k$ (instead of $\alpha_k^2 \leq A_k$).

3.2 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 4 Dual Gradient Method (DGM)

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

$$w_k = \arg\min_{x \in Q} \{\frac{L}{2} \|x - x_k\|_E^2 + \langle g_{\delta,L}(x_k), x - x_k \rangle\} \tag{3.5}$$

5: Compute

$$y_k = \frac{1}{k} \sum_{i=0}^{k} w_i \tag{3.6}$$

6: Compute

$$x_{k+1} = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i \langle g_{\delta,L}(x_i), x - x_i \rangle\} \tag{3.7}$$

7: **end for**

This scheme is nothing else than the Dual Gradient Method (DGM) developed in [11]. 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 [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 behaviors between DGM and FGM:

- In Sections 5 and 6, 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 7, 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 $\Theta(k^{p-1}\delta)$.

4 Optimal choice of the coefficients for a target accuracy ϵ

4.1 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 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 3 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. □

4.2 Lower Complexity bound

If we drop the family of constraint

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i \geq 0 \tag{4.1}$$

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 (4.1), 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 (4.1) and we can only conclude that $k^*(\delta, \theta) \geq k^*_{relax}(\delta, \theta) = \frac{4Ld(x^*)}{\delta\theta^2} - 1$.

5 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:

$$k^*(\delta, \theta) = \min_{\alpha} k \tag{5.1}$$

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 5.4, 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.

5.1 Feasible Policy

This switching policy is feasible for the IGM iff

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_i, \quad \forall i = 0, ..., k \tag{5.2}$$

and

$$\alpha_i \geq 1, \quad \forall i = 0, ..., k. \tag{5.3}$$

When $i \leq m$, the condition (5.2) 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 (5.2) 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 (5.2) is trivially satisfied provided that it is satisfied for $i = m + 1$.
On the other hand, the condition (5.3) 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:

$$k^*_{Switch}(\delta, \theta) = \min_{k \in \mathbb{N}, m \in \mathbb{N}, l} k \tag{5.4}$$

such that

$$k \geq m \tag{5.5}$$

$$\xi(k,\alpha) \geq 0 \tag{5.6}$$

$$\alpha_i = \frac{i+2}{2}, \quad \forall i = 0, ..., m \tag{5.7}$$

$$\alpha_i = l, \quad \forall i = m+1, ..., k \tag{5.8}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{5.9}$$

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

$$\sum_{i=0}^{k} \alpha_i^2 = \sum_{i=0}^{m} \left(\frac{i+2}{2}\right)^2 + \sum_{i=m+1}^{k} l^2 = \frac{1}{24}(m+1)(2m^2+13m+24) + (k-m)l^2$$

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:

$$k^*_{Switch}(\delta, \theta) = \min_{(k,m,l)\in\mathbb{R}^3_+} k \tag{5.10}$$

such that

$$k \geq m. \tag{5.11}$$

$$(k-m)(\theta l - l^2) \geq \frac{1}{24}(m+1)(2m^2+13m+24-6m\theta-24\theta) + \frac{Ld(x^*)}{\delta} \tag{5.12}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{5.13}$$

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 (5.11), (5.12) and (5.13) 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

   $$k \geq m \tag{5.14}$$

$$k \geq m + \frac{Ld(x^*)}{\delta(\theta l - l^2)} + \frac{(m+1)(2m^2+13m+24-6\theta m-24\theta)}{24(\theta l - l^2)} \tag{5.15}$$

$$1 \leq l \leq \frac{m+2}{2}. \tag{5.16}$$

For a given choice of l and m (and dropping the integrality assumption on k), the needed number of iteration is given by

$$k = m + \frac{1}{\theta l - l^2}\max(0, \frac{Ld(x^*)}{\delta} + \frac{1}{24}(m+1)(2m^2+13m+24-6\theta m-24\theta)). \tag{5.17}$$

5.2 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 (5.17), 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 $\theta \geq 2$.

5.3 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 $\theta \geq 2$:

$$\begin{aligned} F(m) &= m + \frac{4}{\theta^2}\left(\frac{Ld(x^*)}{\delta} + \frac{(m+1)}{24}(2m^2+13m+24-6m\theta-24\theta)\right) \\ &\quad \text{when } m \geq \theta - 2 \text{ (i.e. } l = \frac{\theta}{2}) \\ &= m + \frac{4}{(m+2)(2\theta-m-2)}(\frac{Ld(x^*)}{\delta} \\ &\quad + \frac{(m+1)}{24}(2m^2+13m+24-6m\theta-24\theta)) \\ &\quad \text{when } m \leq \theta - 2 \text{ (i.e. } l = \frac{m+2}{2}). \end{aligned}$$

Let us start with the case $\theta \leq 2$.

5.3.1 Case 1 : $\theta \leq 2$

Theorem 4 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 $M(0) = \frac{Ld(x^*)}{\delta(\theta-1)} - 1$. □

5.3.2 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 3 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 7 We assume for the rest of the paper 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 8 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 4 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 5 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 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 7 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 5.
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 5 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 3 and 7, 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 5.2, 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 9 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 8 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 4, 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 7, we conclude that $F'(m) < 0$ for all $m < \overline{m}$. Therefore $\arg\min_{m\geq 0} M(m) = \overline{m}$ since

- For all $m \geq \overline{m} : M(m) = Max(m, F(m)) \geq \overline{m} = M(\overline{m})$
- 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 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 8, 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 10 The FGM that we obtain here in the case $\theta \geq \theta_r$ is not completely equivalent with the original method developed in [8] whose behavior we have studied in the inexact case in [2]. 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 11 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 12 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 $m = \theta - 2 = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$).

5.4 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 4.2, 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 $OptAddPol(\delta, \theta)$:

$$k^*_{add}(\delta, \theta) = \min_{k,\alpha} k \tag{5.18}$$

such that:

$$\xi(k, \alpha) \geq 0 \tag{5.19}$$

$$\alpha_i^2 \leq \sum_{j=0}^{i} \alpha_j, \quad \forall i = 0, ..., k \tag{5.20}$$

$$1 \leq \alpha_i \leq \frac{i+2}{2}, \quad \forall i = 0, ..., k. \tag{5.21}$$

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 3, 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 $\theta \geq \theta_r$.

5.5 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

1. **If $1 \leq \theta \leq 2$**
   - 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, $k^*(\delta, \theta) = k^*_{Switch}(\delta, \theta)$.
2. **If $2 \leq \theta \leq \theta_{\mathbf{r}}$**
   - 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 $\frac{3}{2}$), $k^*(\delta, \theta) \geq \frac{4Ld(x^*)}{\delta\theta^2} - 1$.
3. **If $\theta \geq \theta_{\mathbf{r}}$**
   - 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 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 are the worst-case theoretical bounds obtained in the previous section and not numerical results (whose last section of this paper will be devoted).

6.0.1 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:

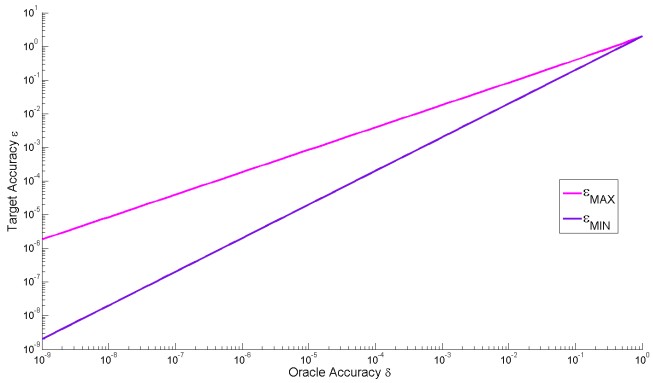

Figure 1: Range of improvement for the IGM with optimal switching strategy.

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.

6.0.2 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.

Let us start with the exact case.

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 $\Theta\left(\frac{LR^2}{\epsilon}\right)$)

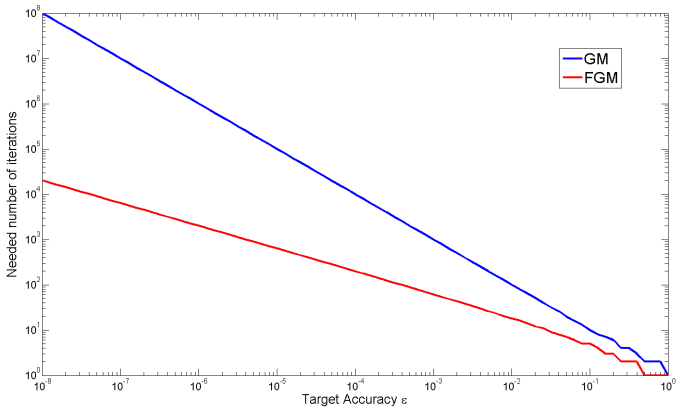

Figure 2: Complexity of the GM and the FGM when used with an exact oracle.

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 13 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 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 iteration 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):

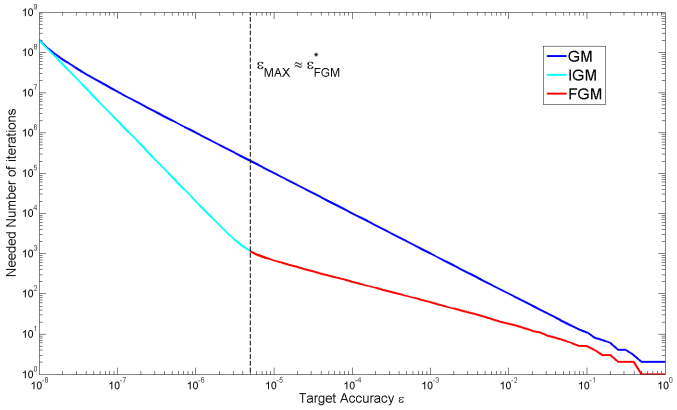

Figure 3: Comparison of the complexities of the GM, the FGM and the IGM when used with an inexact oracle with accuracy $\delta = 5e - 9$.

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:

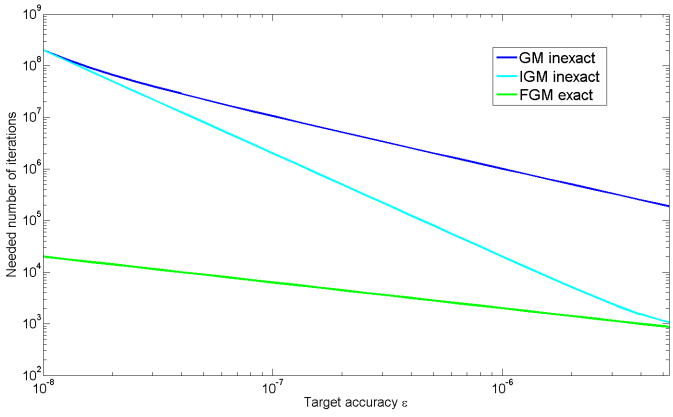

Figure 4: $\delta = 5e - 9$: Complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement of the IGM.

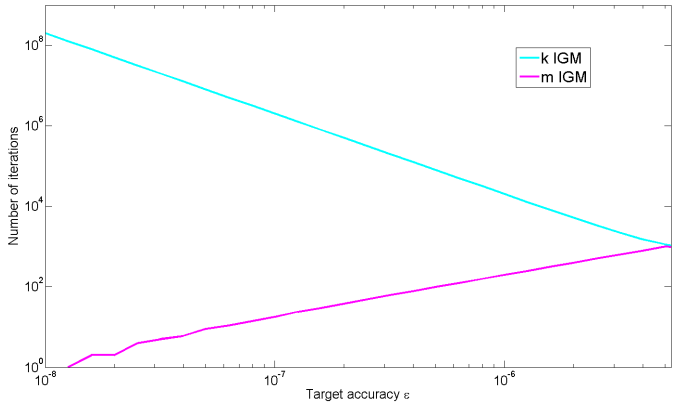

Figure 5: $\delta = 5e-9$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

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

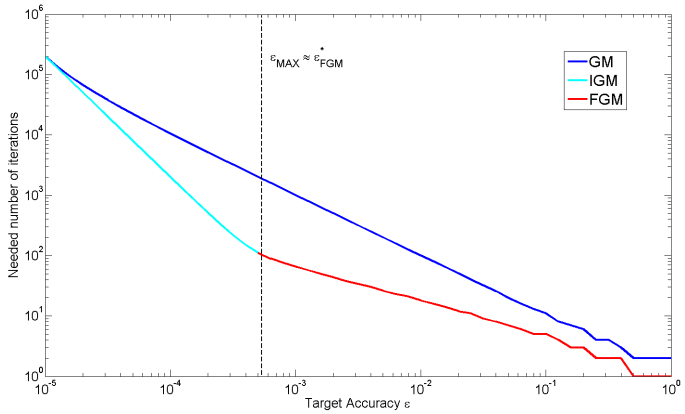

Figure 6: Complexities of the GM, the FGM and the IGM when used with an inexact oracle with accuracy $\delta = 5e-6$.

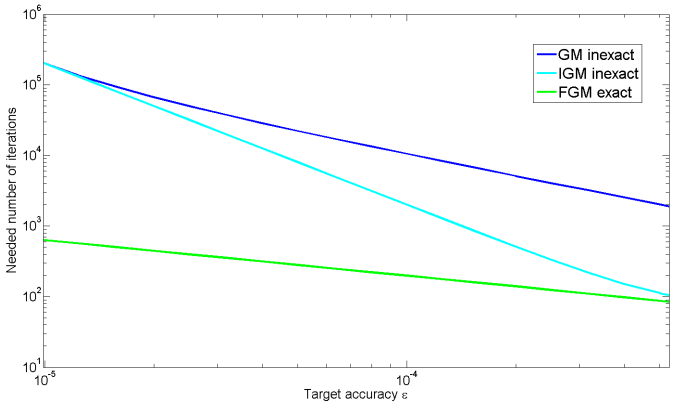

Figure 7: $\delta = 5e-6$: Complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement of the IGM.

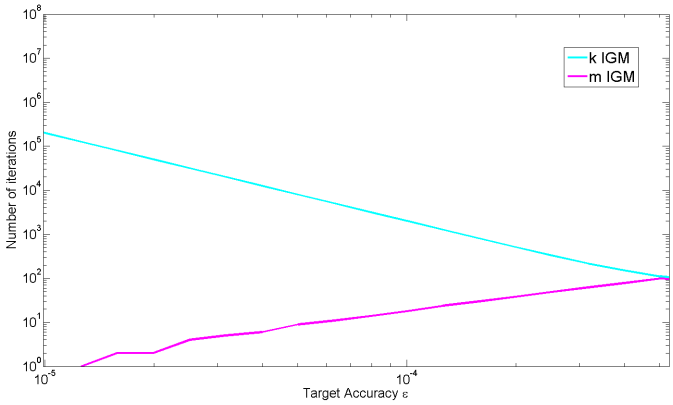

Figure 8: $\delta = 5e-6$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

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

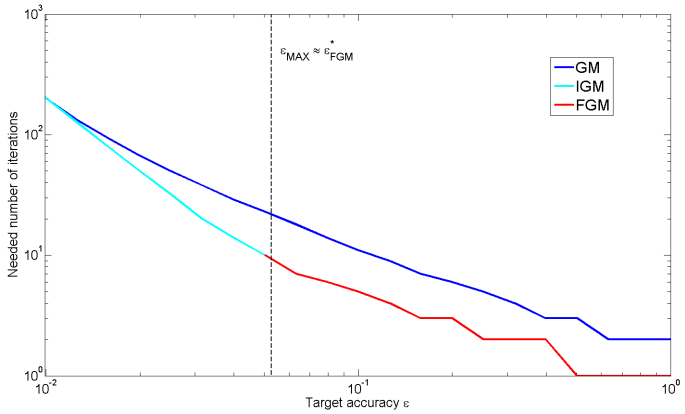

Figure 9: Comparison of the complexities of the GM, the FGM and the IGM when used with an inexact oracle with accuracy $\delta = 5e - 3$.

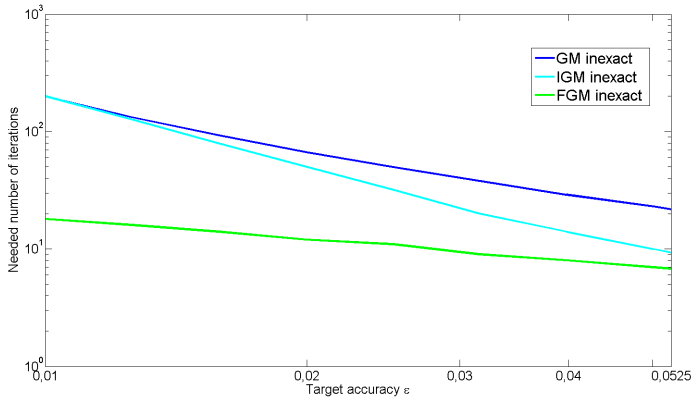

Figure 10: $\delta = 5e - 3$: Comparison of the complexities of the GM in the inexact case, of the IGM in the inexact case and of the FGM in the exact case on the range of improvement.

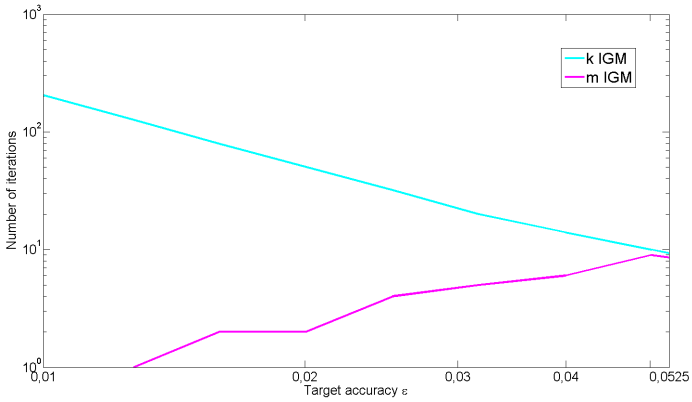

Figure 11: $\delta = 5e-3$: Number of fast-type iterations (m IGM) and total number of iterations (k IGM) for the IGM on its range of improvement.

7 Choice of the coefficients for an intermediate convergence rate

7.1 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 1. 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:

$$f(y_k) - f^* \leq \frac{Ld(x^*)}{\sum_{i=0}^k \alpha_i} + \frac{\sum_{i=0}^k B_i}{\sum_{i=0}^k \alpha_i}\delta. \tag{7.1}$$

Therefore, we would like to find a feasible sequence of coefficients $\{\alpha_i\}_{i\geq 0}$ such that:

1.
$$\sum_{i=0}^k \alpha_i = \Theta(k^p) \tag{7.2}$$

and

2.
$$\frac{\sum_{i=0}^k B_i}{\sum_{i=0}^k \alpha_i} = \Theta(k^{p-1}) \tag{7.3}$$

The condition (7.2) 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:

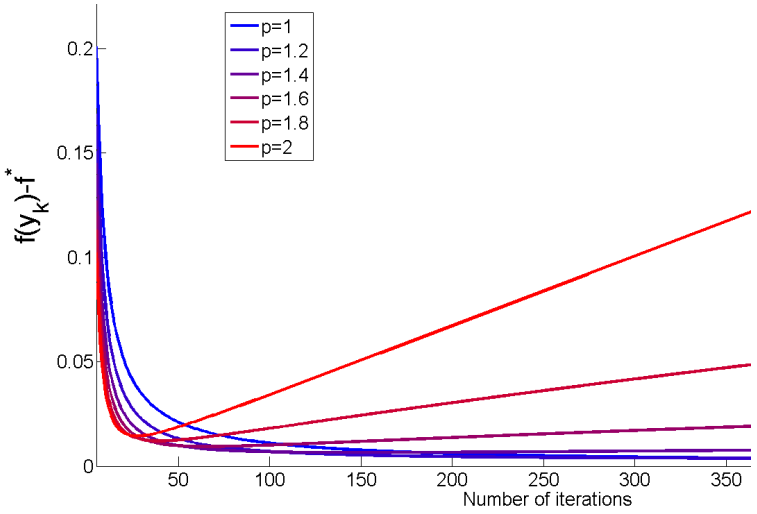

Figure 12: IGM with Power Policy, different convergence rates depending on the choice for p.

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 reach 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.

7.2 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 14 The fact that we need to weaken the upper bound on the convergence in order to be able to analyze the method is one the drawback of the power policy compared to the switching policy.

7.2.1 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 $[0, +\infty[$.

7.2.2 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.

8 Numerical Illustration

Let us finish this paper with a small numerical experiment. Our goal is to observe on a practical example the main results obtained in this paper. We consider the situation of a convex quadratic function on the unit simplex

$$\min_{x\in\Delta_n} \frac{1}{2}x^T Ax \tag{8.1}$$

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 2.2. 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 [3]. 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 [8].

- 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 (8.1) 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 1.1). Three levels of errors are considered: $\delta = 0$, $\delta = 1e-2$ and $\delta = 1e-1$.

8.1 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.

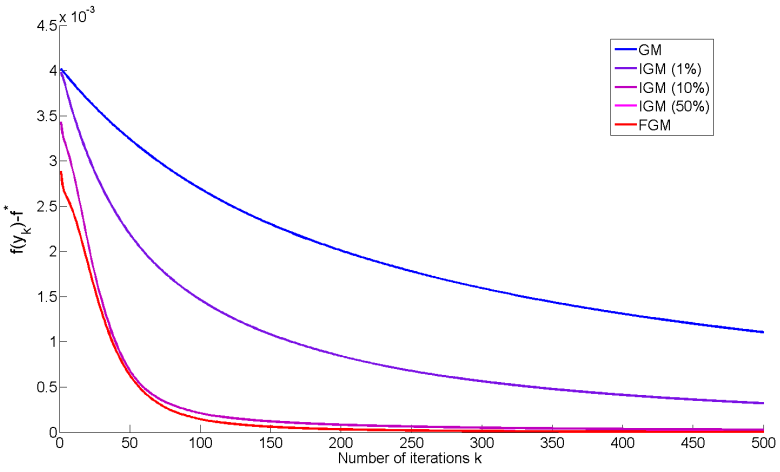

Figure 13: IGM with switching policy: behavior in the exact case

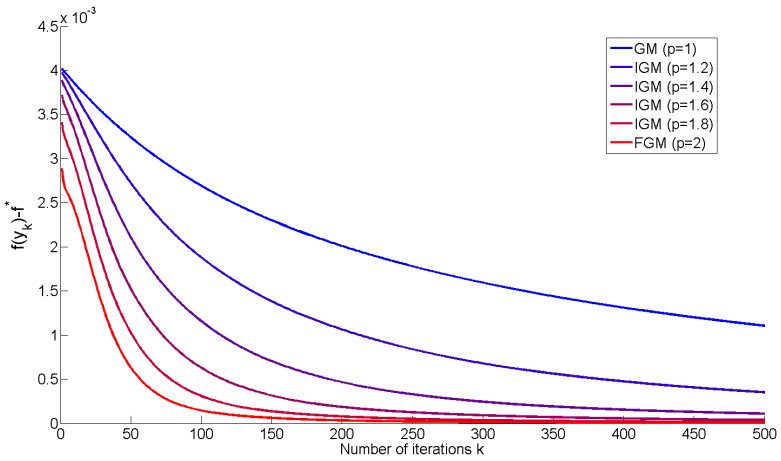

Figure 14: IGM with power policy: behavior in the exact case

Remark 15 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.

8.2 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.

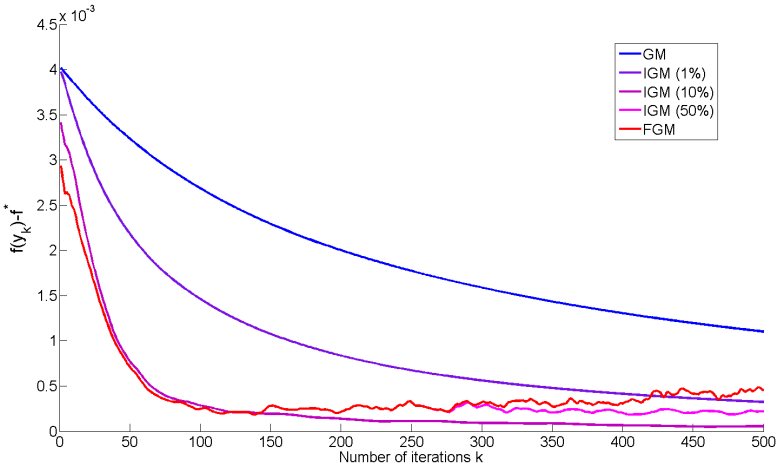

Figure 15: IGM with switching policy: behavior in the inexact case with $\delta = 1e-2$

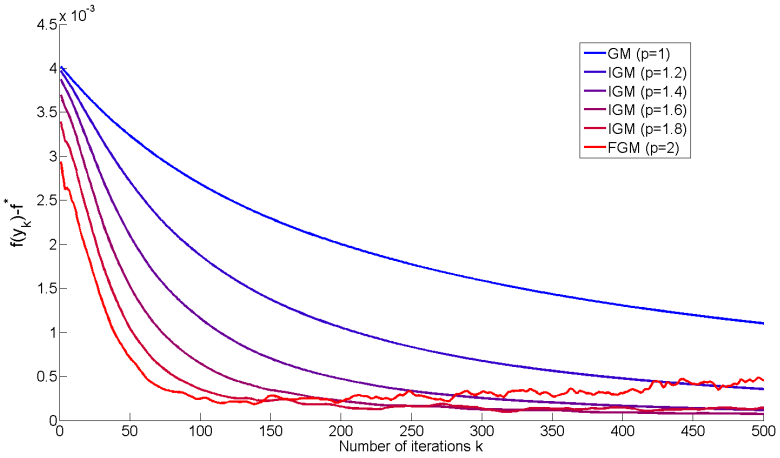

Figure 16: IGM with power policy: behavior in the inexact case with $\delta = 1e-2$

8.3 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:

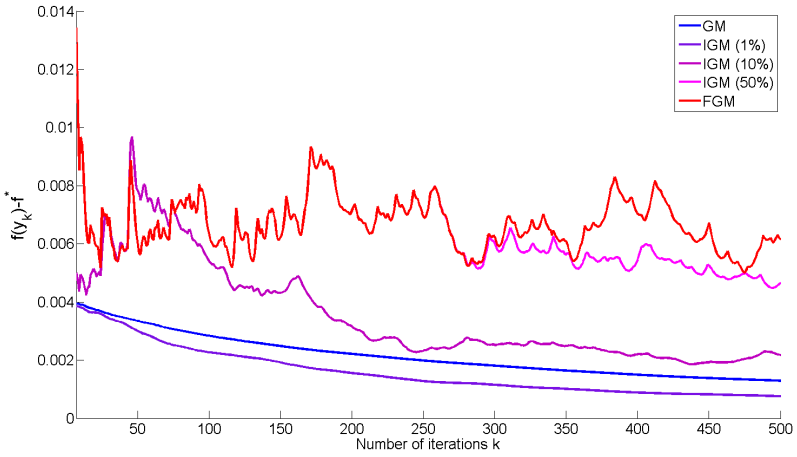

Figure 17: IGM with switching policy: behavior in the inexact case with $\delta = 1e-1$

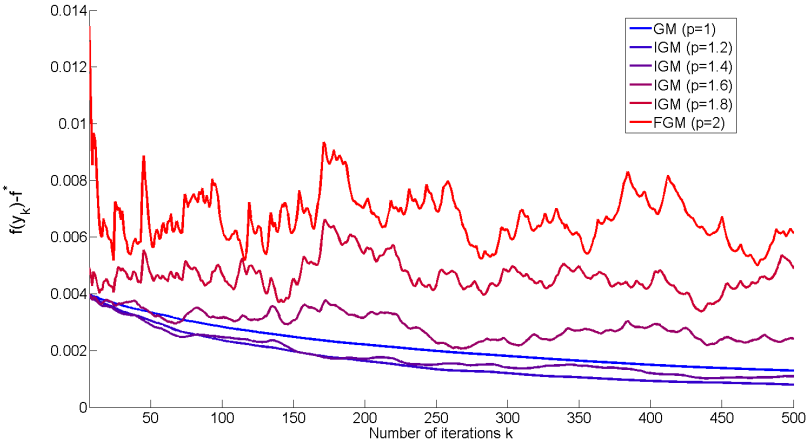

Figure 18: IGM with power policy: behavior in the inexact case with $\delta = 1e-1$

References

[1] O. Devolder, F. Glineur and Y. Nesterov. Double Smoothing Technique for Large-Scale Linearly Constrained Convex Optimization. *SIAM Journal of Optimization*, **22(2)** , (2012)

[2] O. Devolder, F. Glineur and Yu. Nesterov. First-order Methods of Smooth Convex Optimization with Inexact Oracle. *Mathematical Programming, Serie A, Accepted*, (2013).

[3] O. Devolder. Stochastic First Order Methods in Smooth Convex Optimization. *CORE Discussion Paper 2011/70*, (2011).

[4] A. Nemirovskii and D. Yudin.Problem complexity and method efficiency in optimization. *John Wiley* (1983)

[5] 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).

[6] 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).

[7] Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers (2004)

[8] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Serie A*, **103**, 127-152 (2005).

[9] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *Siam Journal of Optimization*, **16**, 235-249 (2005).

[10] Yu. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming A*, **110**, 245-259 (2007).

[11] Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, **76**, (2007)