CORE DISCUSSION PAPER

2013/16

First-Order Methods with Inexact Oracle: the Strongly Convex case

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

April 22, 2013

Abstract

The goal of this paper is to study the effect of inexact first-order information on the first-order methods designed for smooth strongly convex optimization problems. It can be seen as a generalization to the strongly convex case of our previous paper [1].

We introduce the notion of (δ, L, μ)-oracle, that can be seen as an extension of the (δ, L)-oracle (previously introduced in [1]), taking into account strong convexity. We consider different examples of (δ, L, μ)-oracle: strongly convex function with first-order information computed at a shifted point, strongly convex function with approximate gradient and strongly convex max-function with inexact resolution of subproblems.

The core of this paper is devoted to the behavior analysis of three first-order methods, respectively the primal, the dual and the fast gradient method, when used with a (δ, L, μ)-oracle. As in the smooth convex case (studied in [1]), we obtain that the simple gradient methods can be seen as robust but relatively slow, whereas the fast gradient method is faster but more sensitive to oracle errors. However, the strong convexity leads to much faster convergence rates (linear instead of sublinear) for every method and to a reduced sensitivity with respect to oracle errors.

We also prove that the notion of (δ, L, μ)-oracle can be used in order to model exact first-order information but for functions with weaker level of smoothness and different level of convexity. This observation allows us to apply methods, originally designed for smooth strongly convex function, to weakly smooth uniformly convex functions and to derive corresponding performance guarantees.

*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 The (δ, L, μ)-oracle

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^*.

In this paper, E is endowed with an Euclidean norm, defined for a given arbitrary positive definite self-adjoint operator $B : E \to E^*$ by

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

where E^* denotes the dual space of E and $\langle ., . \rangle$, the dual pairing.

1.1 Motivation and definition

Consider $S^{1,1}_{\mu,L}(Q)$, the class of strongly convex functions (with parameter μ) on convex set Q whose gradient is Lipschitz-continuous (with constant L). It is well-known (see [4]) that functions belonging to this class satisfy

$$\frac{\mu}{2} \|x - y\|^2_E \leq f(x) - (f(y) + \langle \nabla f(y), x - y \rangle) \leq \frac{L}{2} \|x - y\|^2_E, \quad \forall x, y \in Q. \tag{1.2}$$

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 $S^{1,1}_{\mu,L}(Q)$ can be characterized by the existence of an oracle returning for each point $y \in Q$ a pair $(f_{L,\mu}(y), g_{L,\mu}(y)) \in \mathbb{R} \times E^*$, necessarily equal to $(f(y), \nabla f(y))$, satisfying

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

Our definition of the (δ, L, μ)-oracle consists in introducing a given amount δ of tolerance in this pair of inequalities:

Definition 1 Let function f be convex on convex set Q. We say that it is equipped with a first-order (δ, L, μ)-oracle if for any $y \in Q$ we can compute a pair $(f_{\delta,L,\mu}(y), g_{\delta,L,\mu}(y)) \in \mathbb{R} \times E^*$ such that

$$\frac{\mu}{2} \|x - y\|^2_E \;\leq\; f(x) - \big(f_{\delta,L,\mu}(y) + \langle g_{\delta,L,\mu}(y), x - y \rangle\big) \leq\; \frac{L}{2} \|x - y\|^2_E + \delta \tag{1.3}$$

for all $x \in Q$ where $\delta \geq 0$ and $L \geq \mu \geq 0$.

This notion of (δ, L, μ)-oracle can be seen as a generalization of the notion of (δ, L)-oracle introduced in [1]. The (δ, L)-oracle has been introduced in order to study the effect of inexact first-order information on the first-order methods designed for an objective function in $F^{1,1}_L(Q) (= S^{1,1}_{0,L}(Q))$. We do the same here but for the first-order methods of $S^{1,1}_{\mu,L}(Q)$.

A function f belongs to $S^{1,1}_{\mu,L}(Q)$ if and only it admits a $(0, L, \mu)$-oracle, namely $(f_{0,L,\mu}(y), g_{0,L,\mu}(y)) = (f(y), \nabla f(y))$. However, the class of functions admitting a (δ, L, μ)-oracle is strictly larger, and also includes both nonsmooth functions and functions that are not strongly convex, as we will see in Section 2.5.

1.2 Properties

The notions of (δ, L, μ) and (δ, L)-oracles are, of course, strongly related:

- A (δ, L, μ)-oracle is also a (δ, L)-oracle.
- A (δ, L)-oracle is a $(\delta, L, 0)$-oracle.

Since a (δ, L, μ)-oracle is also a (δ, L)-oracle, the properties of the (δ, L)-oracle established in Section 2.2 of [1] are also true for a (δ, L, μ)-oracle. In addition, we would like to highlight here two additional properties of a (δ, L, μ)-oracle that will be useful in the rest of this paper:

- If f admits a (δ, L, μ)-oracle, then cf admits a $(c\delta, cL, c\mu)$-oracle for any value of the constant $c > 0$. If f_i admits a (δ_i, L_i, μ_i)-oracle, $i = 1, 2$, then $f_1 + f_2$ admits a $(\delta_1 + \delta_2, L_1 + L_2, \mu_1 + \mu_2)$-oracle.
-

  Theorem 1 If f is endowed with a (δ, L, μ) oracle, we have:

  $$f_{\delta,L,\mu}(\alpha x + (1-\alpha)y) \leq (1-\alpha)f(y) + \alpha f(x) - \frac{\mu}{2}\alpha(1-\alpha)\|y-x\|_E^2$$

  for all $x, y \in E, \alpha \in [0,1]$ and therefore

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

  Proof. Let $x_\alpha = \alpha x + (1-\alpha)y$. We have:

  $$\begin{aligned} f(y) &\geq f_{\delta,L,\mu}(x_\alpha) + \langle g_{\delta,L,\mu}(x_\alpha), y - x_\alpha\rangle + \frac{\mu}{2}\|x_\alpha - y\|_E^2 \\ &= f_{\delta,L,\mu}(x_\alpha) + \alpha\langle g_{\delta,L,\mu}(x_\alpha), y - x\rangle + \frac{\mu}{2}\alpha^2\|y-x\|_E^2 . \end{aligned}$$

  and

  $$\begin{aligned} f(x) &\geq f_{\delta,L,\mu}(x_\alpha) + \langle g_{\delta,L,\mu}(x_\alpha), x - x_\alpha\rangle + \frac{\mu}{2}\|x - x_\alpha\|_E^2 \\ &= f_{\delta,L,\mu}(x_\alpha) + (1-\alpha)\langle g_{\delta,L,\mu}(x_\alpha), x - y\rangle + \frac{\mu}{2}(1-\alpha)^2\|y-x\|_E^2 . \end{aligned}$$

  Adding the first inequality multiplied by $(1-\alpha)$ and the second inequality multiplied by α, we obtain the desired inequality. □

  Therefore if we assume that the function f is endowed with a family of $(\delta, L(\delta), \mu(\delta))$-oracles and that

  1. $\lim_{\delta\to 0}\mu(\delta) = \overline{\mu} > 0$ then we have:

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

     for all $x, y \in E, \alpha \in [0,1]$ and we conclude that f is strongly convex with parameter $\overline{\mu}$
  2. $\lim_{\delta\to 0}\mu(\delta) = 0$ then we have:

     $$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

     for all $x, y \in E, \alpha \in [0,1]$ and we can only conclude that f is convex.

1.3 Paper Structure

In Section 2, we consider different examples of (δ, L, μ) oracle: strongly convex functions with first-order information computed at shifted points, strongly convex functions with approximate gradient, strongly convex max-functions with inexact resolution of subproblems, etc. We prove also that the notion of (δ, L, μ)-oracle can be used in order to model exact first-order information of weakly smooth uniformly convex functions.

Sections 3, 4 and 5 are devoted to the behavior analysis of three first-order methods of smooth strongly convex optimization, respectively the Primal Gradient Method (PGM), the Dual Gradient Method (DGM) and the Fast Gradient Method (FGM), when used with a (δ, L, μ)-oracle. As in the smooth convex case, we obtain that the PGM (and the DGM) can be seen as robust but relatively slow methods, whereas the FGM is faster but more sensitive to oracle errors. However, strong convexity leads to much faster convergence rates (linear instead of sublinear) for every method and to a smaller sensitivity with respect to oracle errors (bounded instead of unbounded accumulation of errors for the FGM).

In Section 6, using the fact that an exact oracle of a weakly smooth uniformly convex function can be seen as a (δ, L, μ)-oracle, we obtain the complexity of our different first-order methods on such kind of objective function. The last section (Section 7) is devoted to the obtainment of lower bounds on the error increase for any first-order method designed for smooth strongly convex functions and used with a (δ, L, μ)-oracle.

2 Examples of (δ, L, μ)-oracle

2.1 Strongly convex function with computation at shifted points

Let function $f \in S^{1,1}_{\mu(f),L(f)}(Q)$ be endowed with an oracle providing at each point $y \in Q$, the exact values of the function and its gradient albeit computed at a shifted point $\hat{y}$ different from y.

1. Since f is strongly convex with parameter $\mu(f)$, we have

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

   Using the convexity of $\|.\|_E^2$, we have $\|x - y\|_E^2 \leq 2\|x - \hat{y}\|_E^2 + 2\|\hat{y} - y\|_E^2$ and therefore

$$f(x) \geq f(\hat{y}) + \langle \nabla f(\hat{y}), x - y \rangle + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \frac{\mu(f)}{4} \|x - y\|_E^2 - \frac{\mu(f)}{2} \|\hat{y} - y\|_E^2. \quad (2.1)$$

2. Since f has a Lipschitz-continuous gradient with constant $L(f)$, we have:

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

   Using the convexity of $\|.\|_E^2$, we have $\|x - \hat{y}\|_E^2 \leq 2\|y - \hat{y}\|_E^2 + 2\|x - y\|_E^2$ and therefore

$$f(x) \leq f(\hat{y}) + \langle \nabla f(\hat{y}), x - y \rangle + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + L(f)\|y - \hat{y}\|_E^2 + L(f)\|x - y\|_E^2. \quad (2.2)$$

Letting $\mu = \frac{\mu(f)}{2}$, $L = 2L(f)$ and $\delta = L(f)\|y - \hat{y}\|_E^2 + \frac{\mu(f)}{2}\|y - \hat{y}\|_E^2$, in view of the equations 2.1 and 2.2, we have that

$$\left(f_{\delta,L,\mu}(y) := f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle - \frac{\mu(f)}{2} \|y - \hat{y}\|_E^2, g_{\delta,L,\mu}(y) := \nabla f(\hat{y})\right)$$

is a (δ, L, μ) oracle for f.

2.2 Functions approximated by a smooth strongly convex function

When a function f can be well approximated by a smooth strongly convex function $\bar{f}$, in the sense that their difference is bounded, the exact values of $\bar{f}$ and its gradient provide a (δ, L, μ)-oracle for f. Indeed, assume that there exists a smooth strongly convex function $\bar{f} \in S^{1,1}_{\mu,L}(Q)$ such that $\bar{f}$ is a δ-lower approximation of f on all Q, i.e.

$$0 \leq f(y) - \bar{f}(y) \leq \delta \quad \forall y \in Q.$$

Using the fact that $\bar{f} \in S^{1,1}_{\mu,L}(Q)$, we obtain

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

(using strong convexity of $\bar{f}$), and

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

(using Lipschitz continuity of $\nabla \bar{f}$), which proves that $(\bar{f}(y), \nabla \bar{f}(y))$ is a (δ, L, μ)-oracle for f.

Finally, note that the above result can be readily extended to the case when the δ-lower approximation $\bar{f}$ is not necessarily smooth and strongly convex but is equipped with an inexact (δ', L, μ) oracle: we can then show that the inexact oracle of $\bar{f}$ also constitutes an inexact $(\delta + \delta', L, \mu)$ oracle for f.

2.3 Strongly convex function with approximate function value and approximate gradient

Let function $f \in S^{1,1}_{\mu(f),L(f)}(Q)$ be endowed with an oracle that provides us at each point $y \in Q$ with an approximate function value $|f(y) - \tilde{f}_y| \leq \Delta_1$ and an approximate gradient $\left\| \nabla f(y) - \tilde{\nabla} f_y \right\|_E^* \leq \Delta_2$.

Let us prove that this very natural definition of approximate first-order information is a particular case of (δ, L, μ) oracle.

As f is strongly convex with parameter $\mu(f)$, we have

$$\begin{aligned} f(x) &\geq f(y) + \langle \tilde{\nabla} f(y), x - y \rangle \\ &\quad + \langle \nabla f(y) - \tilde{\nabla} f(y), x - y \rangle + \frac{\mu(f)}{2} \|x - y\|_E^2 \\ &\geq f(y) + \langle \tilde{\nabla} f(y), x - y \rangle - \Delta_2 \|x - y\|_E + \frac{\mu(f)}{2} \|x - y\|_E^2 \\ &\geq \tilde{f}(y) - \Delta_1 + \langle \tilde{\nabla} f(y), x - y \rangle + \frac{\mu(f)}{4} \|x - y\|_E^2 - \frac{\Delta_2^2}{\mu(f)} \end{aligned}$$

since $\Delta_2 \|x - y\|_E \leq \frac{\Delta_2^2}{\mu(f)} + \frac{\mu(f)}{4} \|x - y\|_E^2$.

As ∇f is Lipschitz-continuous with constant L, we have

$$\begin{aligned} f(x) &\leq f(y) + \langle \tilde{\nabla} f(y), x-y \rangle \\ &\quad + \langle \nabla f(y) - \tilde{\nabla} f(y), x-y \rangle + \frac{L(f)}{2} \|x-y\|_E^2 \\ &\leq f(y) + \langle \tilde{\nabla} f(y), x-y \rangle + \Delta_2 \|x-y\|_E + \frac{L(f)}{2} \|x-y\|_E^2 \\ &\leq \tilde{f}(y) + \Delta_1 + \langle \tilde{\nabla} f(y), x-y \rangle + L(f) \|x-y\|_E^2 + \frac{\Delta_2^2}{2L(f)} \end{aligned}$$

since $\Delta_2 \|x-y\|_E \leq \frac{\Delta_2^2}{2L(f)} + \frac{L(f)}{2} \|x-y\|_E^2$.

We conclude that

$$\left(f_{\delta,L,\mu}(y) = \tilde{f}(y) - \Delta_1 - \frac{\Delta_2^2}{\mu(f)}, g_{\delta,L,\mu}(y) = \tilde{\nabla} f(y) \right)$$

define a (δ, L, μ) oracle for f where $\delta = 2\Delta_1 + \frac{\Delta_2^2}{\mu(f)} + \frac{\Delta_2^2}{2L(f)}$, $\mu = \frac{\mu(f)}{2}$ and $L = 2L(f)$.

In particular, contrarily to the non strongly convex case studied in [1], we see here that boundedness of Q is not needed for an approximate gradient to fit with our definition of inexact oracle.

2.4 Saddle-point functions

Let us now consider objective functions of the form

$$f(x) = \max_{u \in F} \Psi(x,u) = \max_{u \in F} \{G(u) + \langle Au, x \rangle\}$$

where F is a finite-dimensional vector space, endowed with the norm $\|.\|_F$, and $A : F \to E^*$ is a linear operator. We assume that $G : F \to \mathbb{R}$ is

1. Strongly concave with parameter $\mu(G)$ i.e.

$$G(u) \leq G(v) + \langle \nabla G(v), u-v \rangle - \frac{\mu(G)}{2} \|u-v\|_F^2, \quad \forall u, v \in F.$$

2. Smooth with a Lipschitz-continuous gradient with constant $L(G)$

$$G(u) \geq G(v) + \langle \nabla G(v), u-v \rangle - \frac{L(G)}{2} \|u-v\|_F^2, \quad \forall u, v \in F.$$

It is well-known that when $-G \in S^{1,1}_{\mu(G),L(G)}(F)$ then $f \in S^{1,1}_{\mu(f),L(f)}(E)$ where $\mu(f) = \frac{\lambda_{\min}(AA^T)}{L(G)}$ and $L(f) = \frac{\lambda_{\max}(AA^T)}{\mu(G)}$. In particular, the condition numbers of the functions f and G are linked by $Q(f) = \frac{\lambda_{\max}(AA^T)}{\lambda_{\min}(AA^T)} Q(G)$. However, if we want, at a point $z \in E$, to compute the exact first-order information for f, we need to solve the subproblem $\max_{u \in F} \Psi(z,u)$ exactly since $f(z) = \Psi(z, u_z^*)$ and $\nabla f(z) = Au_z^*$ where $u_z^* = \arg\max_{u \in F} \Psi(z,u)$. In practice, we are typically only able to compute an approximate solution $u_z \in F$ of this subproblem. In the following theorem, we give a natural condition under which inexact resolution of the subproblems provides us with a (δ, L, μ)-oracle.

Theorem 2 Assume that G is strongly concave with parameter $\mu(G)$ and smooth with a Lipschitz-continuous gradient with constant $L(G)$. Let z $\in E$ and assume that instead of computing u_z^*, the unique optimal solution of the subproblem $\max_{u \in F} \Psi(z,u)$, we compute $u_z \in F$ such that:

$$\Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi.$$

Then

$$(f_{\delta,L,\mu}(z) = \Psi(z,u_z) - \xi = G(u_z) + \langle Au_z, z\rangle - \xi, g_{\delta,L,\mu}(z) = Au_z)$$

is a (δ, L, μ) oracle for f with $\delta = 3\xi$, $L = \frac{2\lambda_{\max}(AA^T)}{\mu(G)} = 2L(f)$ and $\mu = \frac{\lambda_{\min}(AA^T)}{2L(G)} = \frac{1}{2}\mu(f)$.

Proof. As $\Psi(z,.)$ has a Lipschitz continuous gradient $\nabla_2\Psi(z,u) = \nabla G(u) + A^T z$ with constant $L(G)$, we have (see Theorem 2.1.5. in [4]):

$$(\left\|\nabla G(u_z) + A^T z\right\|_F^*)^2 \leq 2L(G)(\Psi(z,u_z^*) - \Psi(z,u_z)) \leq 2L(G)\xi. \tag{2.3}$$

- The Lipschitz-continuity of ∇G implies (see Lemma 1.2.3 in [4])

$$\begin{aligned} G(u) &\geq G(u_z) + \langle \nabla G(u_z), u - u_z\rangle - \frac{L(G)}{2}\left\|u - u_z\right\|_E^2 \\ &= G(u_z) + \langle -A^T z, u - u_z\rangle + \langle \nabla G(u_z) + A^T z, u - u_z\rangle \\ &\quad - \frac{L(G)}{2}\left\|u - u_z\right\|_F^2. \end{aligned}$$

Therefore:

$$\begin{aligned} f(x) &= \max_{u\in F}\{G(u) + \langle Au, x\rangle\} \\ &\geq \max_{u\in F}[G(u_z) + \langle -A^T z, u - u_z\rangle + \langle \nabla G(u_z) + A^T z, u - u_z\rangle \\ &\quad - \frac{L(G)}{2}\left\|u - u_z\right\|_F^2 + \langle Au, x\rangle] \\ &= G(u_z) + \langle Au_z, z\rangle + \langle Au_z, x - z\rangle + \max_{u\in F}[\langle A(u - u_z), x - z\rangle \\ &\quad + \langle \nabla G(u_z) + A^T z, u - u_z\rangle - \frac{L(G)}{2}\left\|u - u_z\right\|_F^2] \\ &\overset{(2.3)}{\geq} f_{\delta,L,\mu}(z) + \xi + \langle g_{\delta,L,\mu}(z), x - z\rangle + \max_{u\in F}[\langle A(u - u_z), x - z\rangle \\ &\quad - \sqrt{2L(G)\xi}\left\|u - u_z\right\|_F - \frac{L(G)}{2}\left\|u - u_z\right\|_F^2]. \end{aligned}$$

But $\sqrt{2L(G)\xi}\left\|u - u_z\right\|_F \leq \xi + \frac{L(G)}{2}\left\|u - u_z\right\|_F^2$ and therefore:

$$f(x) \geq f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z\rangle + \max_{u\in F}\{\langle A(u - u_z), x - z\rangle - L(G)\left\|u - u_z\right\|_F^2\}.$$

Since

$$\begin{aligned} \max_{u\in F}\{\langle A(u - u_z), x - z\rangle - L(G)\left\|u - u_z\right\|_F^2\} &= \frac{1}{4}\frac{\left\|A^T(x - z)\right\|_{F^*}^2}{L(G)} \\ &\geq \frac{1}{4}\frac{\lambda_{\min}(AA^T)}{L(G)}\left\|x - z\right\|_E^2 \end{aligned}$$

we obtain $f(x) \geq f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z\rangle + \frac{\lambda_{\min}(AA^T)}{4L(G)}\left\|x - z\right\|_E^2$.

- On the other hand, since G is strongly concave with parameter $\mu(G)$, we have:

$$\begin{aligned} G(u) &\leq G(u_z^*) + \langle \nabla G(u_z^*), u - z_z^*\rangle - \frac{\mu(G)}{2}\left\|u - u_z^*\right\|_F^2 \\ &= G(u_z^*) + \langle -A^T z, u - z_z^*\rangle - \frac{\mu(G)}{2}\left\|u - u_z^*\right\|_F^2 \end{aligned}$$

(by definition of u_z^*, we have: $\nabla G(u_z^*) + A^T z = 0$.) Therefore:

$$\begin{aligned} f(x) &= \max_{u\in F}\{G(u) + \langle Au, x\rangle\} \\ &\leq \max_{u\in F}\{G(u_z^*) + \langle -A^T z, u - u_z^*\rangle - \frac{\mu(G)}{2}\|u - u_z^*\|_F^2 + \langle Au, x\rangle\} \\ &= G(u_z^*) + \langle Au_z^*, z\rangle + \langle Au_z^*, x - z\rangle \\ &\quad + \max_{u\in F}\{\langle A(u - u_z^*), x - z\rangle - \frac{\mu(G)}{2}\|u - u_z^*\|_F^2\} \\ &= G(u_z) + (G(u_z^*) - G(u_z)) + \langle Au_z, z\rangle + \langle Au_z, x - z\rangle \\ &\quad + \langle A(u_z^* - u_z), x\rangle + \max_{u\in F}[\langle A(u - u_z), x - z\rangle \\ &\quad + \langle A(u_z - u_z^*), x - z\rangle - \frac{\mu(G)}{2}\|u - u_z^*\|_F^2]. \end{aligned}$$

But $\|u - u_z\|_F^2 \leq 2\|u - u_z^*\|_F^2 + 2\|u_z - u_z^*\|_F^2$ i.e.
$\|u - u_z^*\|_F^2 \geq \frac{1}{2}\|u - u_z\|_F^2 - \|u_z - u_z^*\|_F^2$. Therefore:

$$\begin{aligned} f(x) &\leq G(u_z) + \langle Au_z, z\rangle + \langle Au_z, x - z\rangle + G(u_z^*) - G(u_z) \\ &\quad + \langle A(u_z^* - u_z), z\rangle + \max_{u\in F}[\langle A(u - u_z), x - z\rangle - \frac{\mu(G)}{4}\|u - u_z\|_F^2\} \\ &\quad + \frac{\mu(G)}{2}\|u_z - u_z^*\|_F^2. \end{aligned}$$

Since

1. $\max_{u\in F}\{\langle A(u-u_z), x-z\rangle - \frac{\mu(G)}{4}\|u - u_z\|_F^2\} = \frac{\|A^T(x-z)\|_{F^*}^2}{\mu(G)} \leq \frac{\lambda_{\max}(AA^T)}{\mu(G)}\|x - z\|_E^2$
2. $G(u_z^*) - G(u_z) + \langle A(u_z^* - u_z), z\rangle = \Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi$
3. $\frac{\mu(G)}{2}\|u_z - u_z^*\|_F^2 \leq \Psi(z, u_z^*) - \Psi(z, u_z) \leq \xi$ by strong concavity of $\Psi(z, .)$,

we have:

$$\begin{aligned} f(x) &\leq G(u_z) + \langle Au_z, z\rangle + \langle Au_z, x - z\rangle + \frac{\lambda_{max}(AA^T)}{\mu(G)}\|x - z\|_E^2 + 2\xi \\ &= f_{\delta,L,\mu}(z) + \langle g_{\delta,L,\mu}(z), x - z\rangle + \frac{\lambda_{\max}(AA^T)}{\mu(G)}\|x - z\|_E^2 + 3\xi. \end{aligned}$$

□

2.5 Uniformly convex functions with weaker level of smoothness

Let us show that the notion of (δ, L, μ)-oracle can be also useful for solving problems with exact information but where the objective function is not necessarily strongly convex and ∇f not necessarily Lipschitz-continuous. Let function f be subdifferentiable on Q and for each $y \in Q$, denote by $g(y)$ an arbitrary element of the subdifferential $\partial f(y)$.

We assume that

1. f is uniformly convex on Q with convexity parameters $\rho \geq 2$ and $\kappa > 0$ i.e.:

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

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

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

2. f has an Hölder-continuous (sub)gradient on Q with parameters $\nu \in [0,1]$ an $M < +\infty$ i.e.:
$$\|g(x) - g(y)\|_E^* \leq M \|x - y\|_E^\nu, \quad \forall x, y \in Q.$$
This condition leads to the following inequality:
$$f(x) \leq f(y) + \langle g(y), x - y\rangle + \frac{M}{1+\nu} \|x - y\|_E^{1+\nu}, \quad \forall x, y \in Q.$$

We denote this class of function by:

- $U_{\kappa,M}^{1,\rho,\nu}(Q)$ when the function is also assumed to be differentiable (it is always the case when $\nu > 0$)
- $U_{\kappa,M}^{0,\rho,\nu}(Q)$ when the function can be non-differentiable.

Remark 1
- When $\nu = 0$, the function is typically nonsmooth with bounded variation of the subgradients.
- When $0 < \nu < 1$, the function is weakly-smooth i.e. with a Hölder-continuous gradient.
- When $\nu = 1$, the function is smooth with a Lipschitz-continuous gradient.

In particular when $\nu > 0$, the function is necessarily differentiable and we have
$$U_{\kappa,M}^{1,\rho,\nu}(Q) = U_{\kappa,M}^{0,\rho,\nu}(Q)$$

Remark 2 When $\rho > 1 + \nu$, the class $U_{\kappa,M}^{0,\rho,\nu}(E)$(and therefore also $U_{\kappa,M}^{1,\rho,\nu}(E)$) is empty since
$$\frac{\kappa}{2} t^\rho \geq \frac{M}{1+\nu} t^{1+\nu}$$
for sufficiently large $t \geq 0$.

Remark 3 $U_{\kappa,M}^{1,2,1}(Q) = S_{\kappa,M}^{1,1}(Q)$.

We will prove in this section that functions $f \in U_{\kappa,M}^{0,\rho,\nu}(Q)$, a (δ, L, μ)-oracle is available for any value of $\delta > 0$ i.e. that we can define quantities $(f_{\delta,L,\mu}(y), g_{\delta,L,\mu}(y))$ satisfying inequalities 1.3.

- First, we prove that $f(x) \geq f(y) + \langle g(y), x - y\rangle + \frac{\kappa}{2} \|x - y\|_E^\rho, \quad \forall x, y \in Q$ implies
$$f(x) \geq f(y) + \langle g(y), x - y\rangle + \frac{\mu}{2} \|x - y\|_E^2 - \delta_1, \quad \forall x, y \in Q$$
where $\delta_1 > 0$ is arbitrary and some $\mu > 0$. In order to obtain this implication, we need to find a constant $\mu = \mu(\rho, \kappa, \delta_1)$ such that:
$$\frac{\mu}{2} \|x - y\|_E^2 - \delta_1 \leq \frac{\kappa}{2} \|x - y\|_E^\rho, \quad \forall x, y \in Q.$$
A sufficient condition is $\frac{\mu}{2} t^2 - \delta_1 \leq \frac{\kappa}{2} t^\rho$ for all $t \geq 0$. Therefore we will choose $\mu = \min_{t \geq 0}\{\kappa t^{\rho-2} + 2\delta_1 t^{-2}\}$. The optimal solution of this minimization problem is given by $t^* = \left(\frac{4\delta_1}{\kappa(\rho-2)}\right)^{\frac{1}{\rho}}$ and therefore
$$\mu = \mu(\rho, \kappa, \delta_1) = \rho \left(\frac{1}{\rho - 2}\right)^{\frac{\rho-2}{\rho}} \kappa^{\frac{2}{\rho}} \delta_1^{\frac{\rho-2}{\rho}} 2^{1-\frac{4}{\rho}}.$$
In particular, when $\rho = 2$, we obtain $\mu = \kappa$.

- Second, in the subsection 2.3.c. in [1], we have proved that $f(x) \leq f(y) + \langle g(y), x - y\rangle + \frac{M}{1+\nu} \|x - y\|_E^{1+\nu}$ implies

$$f(x) \leq f(y) + \langle g(y), x - y\rangle + \frac{L}{2} \|x - y\|_E^2 + \delta_2$$

where δ_2 is arbitrary and $L = M \left(\frac{M}{2\delta_2} \frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}$. In particular when $\nu = 1$, we obtain $L = M$.

We obtain the following theorem:

Theorem 3 Assume that $f \in U_{\kappa,M}^{0,\rho,\nu}(Q)$. Let $0 < \delta_1$ and $0 < \delta_2$ be arbitrary constants and define $\delta = \delta_1 + \delta_2$, $\mu = \rho \left(\frac{1}{\rho-2}\right)^{\frac{\rho-2}{\rho}} \kappa^{\frac{2}{\rho}} \delta_1^{\frac{\rho-2}{\rho}} 2^{1-\frac{4}{\rho}}$ and $L = M \left(\frac{M}{2\delta_2} \frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}$. Then

$$(f_{\delta,L,\mu}(y) = f(y) - \delta_1, g_{\delta,L,\mu}(y) = g(y) \in \partial f(y))$$

defines a (δ, L, μ) oracle for f.

3 Primal gradient method with (δ, L, μ)-oracle

Let us now study the behavior of first-order methods, initially developed for smooth strongly convex problems, but used here with a (δ, L, μ)-oracle.

We start with the Primal Gradient Method. One important property of this method is that it does not use the strongly convex parameter μ explicitly in the scheme. The Primal Gradient Method for strongly convex problems looks exactly the same as in the convex case.

Therefore, the Primal Gradient Method when used with a (δ, L, μ) oracle looks exactly the same that when used with a (δ, L)-oracle in [1].

Algorithm 1 Primal Gradient Method (PGM) with (δ, L, μ) oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute $x_{k+1} = \arg\min_{x \in Q}\{\langle \nabla g_{\delta,L,\mu}(x_k), x - x_k\rangle + \frac{L}{2} \|x - x_k\|_E^2\}$
5: **end for**

Even if the scheme is the same, the fact that we use a (δ, L, μ)-oracle instead of a (δ, L)-oracle can accelerate significantly the convergence rate:

Theorem 4 Assume that f is endowed with a (δ, L, μ)-oracle with $\mu > 0$, then the sequence $y_k = \arg\min_{x_1,\ldots,x_k} f(x_i)$, generated by the Primal Gradient Method satisfies

$$f(y_k) - f^* \leq \frac{LR^2}{2} \exp\left(-k\frac{\mu}{L}\right) + \delta$$

where $R = \|x - x_0\|_E$.

Proof. Denote $r_k = \|x_k - x^*\|_E$ and $f_k = f_{\delta,L,\mu}(x_k)$, $g_k = g_{\delta,L,\mu}(x_k)$. We have

$$r_{k+1}^2 = \|x_{k+1} - x^*\|_E^2 = r_k^2 + 2\langle B(x_{k+1} - x_k), x_{k+1} - x^*\rangle - \|x_{k+1} - x_k\|_E^2 \qquad (3.1)$$

Using the optimality condition of the problem defining x_{k+1}:

$$\langle g_k + LB(x_{k+1} - x_k), x - x_{k+1} \rangle \geq 0 \quad \forall x \in Q$$

we have

$$\langle B(x_{k+1} - x_k), x_{k+1} - x^* \rangle \leq \frac{1}{L} \langle g_k, x^* - x_{k+1} \rangle.$$

We obtain

$$\begin{aligned}
r_{k+1}^2 &\leq r_k^2 + \frac{2}{L}\langle g_k, x^* - x_{k+1}\rangle - \|x_{k+1} - x_k\|_E^2 \\
&= r_k^2 + \frac{2}{L}\langle g_k, x^* - x_k\rangle - \frac{2}{L}\left[\langle g_k, x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_E^2\right] \\
&\overset{(1.3)}{\leq} r_k^2 + \frac{2}{L}\langle g_k, x^* - x_k\rangle - \frac{2}{L}\left[f(x_{k+1}) - f_k - \delta\right] \\
&\overset{(1.3)}{\leq} r_k^2 + \frac{2}{L}\left[f(x^*) - f_k - \frac{\mu}{2}\|x_k - x^*\|_E^2\right] - \frac{2}{L}\left[f(x_{k+1}) - f_k - \delta\right] \\
&= \left(1 - \frac{\mu}{L}\right) r_{k+1}^2 + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
r_{k+1}^2 &\leq \left(1 - \frac{\mu}{L}\right) r_k^2 + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right] \\
&\leq \left(1 - \frac{\mu}{L}\right)\left(\left(1 - \frac{\mu}{L}\right) r_{k-1}^2 + \frac{2}{L}\left[f(x^*) - f(x_k) + \delta\right]\right) \\
&\quad + \frac{2}{L}\left[f(x^*) - f(x_{k+1}) + \delta\right] \\
&\leq \left(1 - \frac{\mu}{L}\right)^k r_0^2 + \frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i \left(f(x^*) - f(x_{k+1-i}) + \delta\right).
\end{aligned}$$

and we obtain

$$\frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i \left(f(x_{k+1-i}) - f(x^*)\right) \leq \left(1 - \frac{\mu}{L}\right)^{k+1} r_0^2 + \frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i \delta.$$

Therefore, using the definition of y_{k+1} and the fact that $\frac{2}{L}\sum_{i=0}^{k}\left(1 - \frac{\mu}{L}\right)^i = \frac{2}{\mu}\left(1 - (1 - \frac{\mu}{L})^{k+1}\right)$ we conclude that

$$\begin{aligned}
f(y_{k+1}) - f^* &\leq \frac{\mu}{2}\frac{\left(1 - \frac{\mu}{L}\right)^{k+1}}{1 - \left(1 - \frac{\mu}{L}\right)^{k+1}} r_0^2 + \delta \\
&\leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{k+1} r_0^2 + \delta \\
&\leq \frac{L r_0^2}{2}\exp\left(-(k+1)\frac{\mu}{L}\right) + \delta.
\end{aligned}$$

□

Remark 4 When $\delta = 0$, we retrieve the well-known behavior of the Primal Gradient Method in the strongly convex case, with a complexity of order $\Theta\left(\frac{L}{\mu}\ln\left(\frac{LR^2}{\epsilon}\right)\right)$.

Remark 5 As a (δ, L, μ) oracle is also a (δ, L)-oracle and as the parameter μ is not used during the scheme, the upper-bound

$$f(y_k) - f^* \leq \frac{LR^2}{2k} + \delta \tag{3.2}$$

obtained in [1], is still available. Therefore, the sequence y_k generated by the primal gradient method actually satisfies

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

In conclusion, when we apply the primal gradient method to a function endowed with a (δ, L, μ) oracle, there is no error accumulation, and the upper bound for the objective function accuracy decreases with k and asymptotically tends to δ. If we want an accuracy of ϵ for the objective function, we need to perform a number of iterations k such that

$$k = \min\left(\Theta\left(\frac{LR^2}{\epsilon}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$$

with an oracle accuracy $\delta = \Theta(\epsilon)$. As in the non strongly convex case, studied in [1], the PGM does not suffer from errors accumulation.

4 Dual gradient method with (δ, L, μ)-oracle

Let us now consider the Dual Gradient Method. This method has been introduced in [6] for smooth convex problems with exact oracle. In [1], we have studied its behavior when used with a (δ, L)-oracle.

In the strongly convex case, it is necessary to modify slightly the method in order to take advantage of the strong convexity. We propose here such modification and study the behavior of this modified method when used with a (δ, L, μ) oracle.

Let $\{\alpha_k\}_{k\geq 0}$ be a sequence of positive reals such that:

$$\alpha_0 = \frac{L}{L-\mu} \tag{4.1}$$

$$(L-\mu)\alpha_{k+1} = A_k\mu + L \tag{4.2}$$

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

Algorithm 2 Dual Gradient Method (DGM) with (δ, L, μ) oracle

1: Choose $x_0 \in Q$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute

$$w_k = \arg\min_{x\in Q}\{\langle g_{\delta,L,\mu}(x_k), x - x_k\rangle + \frac{L}{2}\|x - x_k\|_E^2\} \tag{4.3}$$

5: Compute

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

6: **end for**

Lemma 1 For any $k \geq 0$ we have

$$\sum_{i=0}^{k} \alpha_i [f(w_i) - f^*] \quad \leq \quad \tfrac{L}{2} \|x_0 - x^*\|_E^2 + \sum_{i=0}^{k} \alpha_i \delta \tag{4.4}$$

Proof. For $k \geq 0$, denote $f_k = f_{\delta,L,\mu}(x_k)$, $g_k = g_{\delta,L,\mu}(x_k)$, and $\psi_k^* = \min_{x \in Q} \psi_k(x)$ where

$$\psi_k(x) \quad = \quad \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, x - x_i \rangle + \tfrac{\mu}{2} \|x - x_i\|_E^2] + \tfrac{L}{2} \|x - x_0\|_E^2.$$

In view of the first inequality in (1.3), we have for all $x \in Q$

$$\psi_k^* \quad \leq \quad \psi_k(x) \; \leq \; \textstyle\sum_{i=0}^{k} \alpha_i f(x) + \frac{L}{2} \|x - x_0\|_E^2 . \tag{4.5}$$

Let us prove that $\psi_k^* \geq \sum\limits_{i=0}^{k} \alpha_i [f(w_i) - \delta], \quad \forall k \geq 0.$
Indeed, this inequality is valid for $k = 0$:

$$\begin{aligned}
\alpha_0 f(w_0) &\overset{(1.3)}{\leq} \alpha_0 [f_0 + \langle g_0, w_0 - x_0 \rangle + \frac{L}{2} \|w_0 - x_0\|_E^2 + \delta] \\
&\overset{(4.3)}{=} \min_{y \in Q} \alpha_0 [f_0 + \langle g_0, y - x_0 \rangle + \frac{L}{2} \|y - x_0\|_E^2] + \alpha_0 \delta \\
&\leq \min_{y \in Q} \{\alpha_0 [f_0 + \langle g_0, y - x_0 \rangle + \frac{\mu}{2} \|y - x_0\|_E^2] + \frac{L}{2} \|y - x_0\|_E^2\} + \alpha_0 \delta \\
&= \psi_0^* + \alpha_0 \delta.
\end{aligned}$$

Assume it is valid for some $k \geq 1$. Since $\Psi_k(x)$ is strongly convex with parameter $\sum_{i=0}^{k} \alpha_i \mu + L = A_k \mu + L$, we have:

$$\psi_k(x) \quad \geq \quad \psi_k^* + \tfrac{A_k \mu + L}{2} \|x - x_{k+1}\|_E^2, \quad x \in Q$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &= \min_{x \in Q} \Big\{ \psi_k(x) + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \Big\} \\
&\geq \psi_k^* + \min_{x \in Q} \{ \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\
&\quad + \frac{A_k \mu + L}{2} \|x - x_{k+1}\|_E^2 \} \\
&\geq \psi_k^* + \alpha_{k+1} \min_{x \in Q} \{ f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{L}{2} \|x - x_{k+1}\|_E^2 \}
\end{aligned}$$

since $\alpha_{k+1} \mu + A_k \mu + L = L \alpha_{k+1}$.
And we obtain finally :

$$\psi_{k+1}^* \overset{(4.3),(1.3)}{\geq} \psi_k^* + \alpha_{k+1} (f(w_{k+1}) - \delta).$$

Hence, using our inductive assumption, we have proved that $\psi_k^* \geq \sum\limits_{i=0}^{k} \alpha_i [f(w_i) - \delta]$ for all $k \geq 0$. To conclude, we combine this fact with inequality (4.5) for $x = x^*$. □

Defining now the approximate solution as $y_k = \arg\min_{i=0,\ldots,k} f(w_i)$ or $y_k = \frac{\sum_{i=0}^{k} \alpha_i w_i}{\sum_{i=0}^{k} \alpha_i}$ we obtain:

$$f(y_k) - f^* \leq \frac{L \|x_0 - x^*\|_E^2}{2 \sum_{i=0}^{k} \alpha_i} + \delta = \frac{L R^2}{2 A_k} + \delta. \tag{4.6}$$

It remains therefore to obtain a lower bound for A_k:

Lemma 2 The sequence $\{A_k\}_{k\geq 0}$ defined by the recurrence (4.1) and (4.2) satisfies

$$A_k = \sum_{i=1}^{k+1} \left(\frac{L}{L-\mu}\right)^i$$

and therefore

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

Proof. We have

$$(L-\mu)\alpha_{k+1} = A_k\mu + L \Leftrightarrow (L-\mu)A_{k+1} = L(A_k+1)$$

and therefore

$$A_{k+1} = \frac{L}{L-\mu}(A_k+1).$$

As $A_0 = \alpha_0 = \frac{L}{L-\mu}$, we conclude that

$$A_k = \sum_{i=1}^{k+1} \left(\frac{L}{L-\mu}\right)^i.$$

□

We obtain finally the following theorem:

Theorem 5 The dual gradient method applied to a function f endowed with a (δ, L, μ)-oracle generates a sequence $\{w_k\}_{k\geq 0}$ such that $y_k = \arg\min_{i=0,\ldots,k} f(w_i)$ and $y_k = \frac{\sum_{i=0}^k \alpha_i w_i}{\sum_{i=0}^k \alpha_i}$ satisfy

$$f(y_k) - f^* \leq \frac{LR^2}{2(k+1)} + \delta, \text{ if } \mu = 0$$

and

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

if $\mu > 0$.

Remark 6 When $\mu = 0$, we have $\alpha_i = 1$ $\forall i \geq 0$ and this method corresponds to the dual gradient method introduced in [6] and for which the behavior when used with a $(\delta, L) = (\delta, L, 0)$ has been already established in [1].

Remark 7 In the case $\mu > 0$, the sequence $A_k(\mu) = A_k$ satisfies the recurrence

$$A_{k+1}(\mu) = \frac{L}{L-\mu}A_k(\mu) + \frac{L}{L-\mu}$$

and in the case $\mu = 0$, the sequence $A_k(0)$ satisfies the recurrence

$$A_{k+1}(0) = A_k(0) + 1.$$

Clearly $A_k(\mu) \geq A_k(0)$ for all $k \geq 1$ and

$$f(y_k) - f^* \leq \frac{LR^2}{2A_k(\mu)} + \delta \leq \frac{LR^2}{2A_k(0)} + \delta.$$

Therefore the upper-bound

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

is also available in the case $\mu > 0$ and we have:

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

Remark 8 When $\delta = 0$, the availability of a $(0, L, \mu)$ oracle for a function f means simply that $f \in S^{1,1}_{\mu,L}(Q)$. To the best of our knowledge, it is the first time that the dual gradient method is adapted to the strongly convex case.

Since we obtain the same convergence results for both primal and dual gradient methods, we will refer to both as Gradient Methods (GM) in the rest of this paper.

5 Fast gradient method with (δ, L, μ)-oracle

The fast gradient method (at least the version that we consider in this paper) has been introduced in [5]. In [1], we have studied the behavior of this scheme when used with a (δ, L)-oracle instead of the exact one. In this section, we adapt this fast-gradient method to the strongly convex case and we apply this scheme to a convex function f endowed with a (δ, L, μ)-oracle.

5.1 The method

Let $\{\alpha_k\}_{k\geq 0}$ be a sequence of reals such that

$$L + \mu A_k = \frac{L\alpha_{k+1}^2}{A_{k+1}}, \quad \alpha_0 = 1 \tag{5.1}$$

where $A_k = \sum_{i=0}^{k} \alpha_i$. Define $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$, $k \geq 0$. The condition on the sequence $\{\alpha_k\}_{k=0}^{\infty}$ is equivalent with

$$\frac{L + \mu A_k}{A_{k+1}} = L\tau_k^2. \tag{5.2}$$

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

$$d(x_0) = 0, \quad d(x) \geq \frac{1}{2}\|x - x_0\|_E^2, \quad \forall x \in Q. \tag{5.3}$$

For a particular choice of the prox-function, the FGM used with a (δ, L, μ)-oracle looks as follows

Algorithm 3 Fast Gradient Method (FGM) with (δ, L, μ) oracle

1: Choose $x_0 = \min_{x \in Q} d(x)$
2: **for** k = 0 : ... **do**
3: Obtain $(f_{\delta,L,\mu}(x_k), g_{\delta,L,\mu}(x_k))$.
4: Compute

$$y_k = \arg\min_{x \in Q} \{\langle g_{\delta,L,\mu}(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_E^2\} \tag{5.4}$$

5: Compute $z_k = \arg\min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [\langle g_{\delta,L,\mu}(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2]\}$
6: Define $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
7: **end for**

5.2 Convergence rate

Denote

$$\psi_k^* = \min_{x \in Q} \{Ld(x) + \sum_{i=0}^{k} \alpha_i [f_{\delta,L,\mu}(x_i) + \langle g_{\delta,L,\mu}(x_i), x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2]\}.$$

Lemma 3 For all $k \geq 0$, we have $A_k f(y_k) \leq \psi_k^* + E_k$ with $E_k = \sum_{i=0}^{k} A_i \delta$.

Proof. Denote $f_k = f_{\delta,L,\mu}(x_k)$, and $g_k = g_{\delta,L,\mu}(x_k)$. For $k = 0$, we have

$$\begin{aligned} \psi_0^* &= \min_{x \in Q} \left\{ Ld(x) + \alpha_0 [f_0 + \langle g_0, x - x_0 \rangle + \tfrac{\mu}{2} \|x - x_0\|_E^2] \right\} \\ &\overset{(5.3)}{\geq} \min_{x \in Q} \left\{ f_0 + \langle g_0, x - x_0 \rangle + \tfrac{L}{2} \|x - x_0\|^2 \right\} \overset{(1.3)}{\geq} [f(y_0) - \delta]. \end{aligned}$$

since $\alpha_0 = 1$.

Assume now that the statement of the lemma is true for some $k \geq 0$. Optimality condition for the optimization problem defining z_k implies

$$\langle \nabla Ld(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i g_i + \sum_{i=0}^{k} \alpha_i \mu B(z_k - x_i), x - z_k \rangle \quad \geq \quad 0, \quad \forall x \in Q.$$

Hence, in view of strong convexity of d,

$$\begin{aligned} Ld(x) &\geq Ld(z_k) + \langle L \nabla d(z_k), x - z_k \rangle + \frac{L}{2} \|x - z_k\|_E^2 \\ &\geq Ld(z_k) + \sum_{i=0}^{k} \alpha_i \langle g_i, z_k - x \rangle \\ &\quad + \sum_{i=0}^{k} \alpha_i \mu \langle B(z_k - x_i), z_k - x \rangle + \frac{L}{2} \|x - z_k\|_E^2 . \end{aligned}$$

Thus, we have for all $x \in Q$:

$$\begin{aligned} & Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2] \\ \geq \; & Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i \rangle] + \frac{L}{2} \|x - z_k\|_E^2 \\ & + \sum_{i=0}^{k} \alpha_i \mu \langle B(z_k - x_i), z_k - x \rangle + \sum_{i=0}^{k} \frac{\alpha_i \mu}{2} \|x - x_i\|_E^2 \\ & + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2]. \end{aligned}$$

But:

$$\langle B(z_k - x_i), z_k - x \rangle = \frac{1}{2} \|z_k - x_i\|_E^2 + \frac{1}{2} \|z_k - x\|_E^2 - \frac{1}{2} \|x - x_i\|_E^2$$

and we obtain:

$$\begin{aligned} & Ld(x) + \sum_{i=0}^{k+1} \alpha_i [f_i + \langle g_i, x - x_i \rangle + \frac{\mu}{2} \|x - x_i\|_E^2] \\ & \geq Ld(z_k) + \sum_{i=0}^{k} \alpha_i [f_i + \langle g_i, z_k - x_i \rangle + \frac{\mu}{2} \|z_k - x_i\|_E^2] \\ & + \frac{L + A_k \mu}{2} \|z_k - x\|_E^2 + + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2]. \end{aligned}$$

which implies:

$$\begin{aligned} \psi_{k+1}^* \quad \geq \quad & \psi_k^* + \min_{x \in Q} \{ \tfrac{L + \mu A_k}{2} \|x - z_k\|_E^2 + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle \\ & + \tfrac{\mu}{2} \|x - x_{k+1}\|_E^2] \}. \end{aligned}$$

On the other hand, using our recurrence assumption $A_k f(y_k) \leq \psi_k^* + E_k$, we have

$$\begin{aligned} & \psi_k^* + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\ \geq \quad & A_k f(y_k) - E_k + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\ \overset{(1.3)}{\geq} \quad & A_k [f_{k+1} + \langle g_{k+1}, y_k - x_{k+1} \rangle + \frac{\mu}{2} \|y_k - x_{k+1}\|_E^2] - E_k \\ & + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\ = \quad & A_{k+1} f_{k+1} + \langle g_{k+1}, A_k (y_k - x_{k+1}) + \alpha_{k+1} (x - x_{k+1}) \rangle \\ & - E_k + \frac{A_k \mu}{2} \|y_k - x_{k+1}\|_E^2 + \frac{\alpha_{k+1} \mu}{2} \|x - x_{k+1}\|_E^2 . \end{aligned}$$

Taking into account that

$$\begin{aligned} & A_k (y_k - x_{k+1}) + \alpha_{k+1} (x - x_{k+1}) \\ = \; & A_k \tau_k (y_k - z_k) + \alpha_{k+1} x - \alpha_{k+1} \tau_k z_k - \alpha_{k+1} (1 - \tau_k) y_k \; = \; \alpha_{k+1} (x - z_k), \end{aligned}$$

we obtain

$$\begin{aligned} & \psi_k^* + \alpha_{k+1} [f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle + \frac{\mu}{2} \|x - x_{k+1}\|_E^2] \\ \geq \quad & A_{k+1} f_{k+1} + \alpha_{k+1} \langle g_{k+1}, x - z_k \rangle - E_k. \end{aligned}$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &\geq A_{k+1}f_{k+1} - E_k + \min_{x\in Q}\{\tfrac{L+\mu A_k}{2}\|x-z_k\|_E^2 + \alpha_{k+1}\langle g_{k+1}, x-z_k\rangle\} \\
&= A_{k+1}\left[f_{k+1} + \min_{x\in Q}\{\tfrac{L+\mu A_k}{2A_{k+1}}\|x-z_k\|_E^2 + \tau_k\langle g_{k+1}, x-z_k\rangle\}\right] - E_k \\
&\overset{(5.2)}{=} A_{k+1}\left[f_{k+1} + \min_{x\in Q}\{\tfrac{\tau_k^2 L}{2}\|x-z_k\|_E^2 + \tau_k\langle g_{k+1}, x-z_k\rangle\}\right] - E_k.
\end{aligned}$$

For $x \in Q$, define $y = \tau_k x + (1-\tau_k)y_k$. Since $y - x_{k+1} = \tau_k(x - z_k)$, we obtain

$$\begin{aligned}
&\min_{x\in Q}\left\{\tfrac{\tau_k^2 L}{2}\|x-z_k\|_E^2 + \tau_k\langle g_{k+1}, x-z_k\rangle\right\} \\
&= \min_{y}\left\{\tfrac{L}{2}\|y-x_{k+1}\|_E^2 + \langle g_{k+1}, y-x_{k+1}\rangle : \ y \in \tau_k Q + (1-\tau_k)y_k\right\} \qquad (5.5) \\
&\geq \min_{y\in Q}\left\{\tfrac{L}{2}\|y-x_{k+1}\|_E^2 + \langle g_{k+1}, y-x_{k+1}\rangle\right\}.
\end{aligned}$$

Therefore,we have

$$\begin{aligned}
\psi_{k+1}^* &\geq A_{k+1}\left[f_{k+1} + \min_{y\in Q}\{\tfrac{L}{2}\|y-x_{k+1}\|^2 + \langle g_{k+1}, y-x_{k+1}\rangle\}\right] - E_k \\
&\overset{(5.4),(1.3)}{\geq} A_{k+1}f(y_{k+1}) - E_k - A_{k+1}\delta,
\end{aligned}$$

and we get $A_{k+1}f(y_{k+1}) \leq \psi_{k+1} + E_{k+1}$ with $E_{k+1} = E_k + A_{k+1}\delta$. □

As a direct consequence of this lemma, we obtain

Theorem 6 For all $k \geq 0$, we have $f(y_k) - f^* \leq \frac{1}{A_k}\left(Ld(x^*) + \sum_{i=0}^k A_i\delta\right)$.

Proof. Denote $f_i = f_{\delta,L,\mu}(x_i)$, and $g_i = g_{\delta,L,\mu}(x_i)$. Then

$$\begin{aligned}
\psi_k^* &= \min_{x\in Q}\left\{Ld(x) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x-x_i\rangle + \tfrac{\mu}{2}\|x-x_i\|_E^2]\right\} \\
&\leq Ld(x^*) + \sum_{i=0}^k \alpha_i[f_i + \langle g_i, x^*-x_i\rangle + \tfrac{\mu}{2}\|x^*-x_i\|_E^2] \\
&\overset{(1.3)}{\leq} Ld(x^*) + A_k f(x^*).
\end{aligned}$$

The proof now simply follows from the recurrence established in Lemma 3. □

It remains to estimate A_k and $\frac{\sum_{i=0}^k A_i}{A_k}$. More precisely, in order to obtain an explicit upper-bound for the convergence rate of this method, we need

1. A lower-bound for A_k
2. An upper-bound for $\frac{\sum_{i=0}^k A_i}{A_k}$.

Concerning the lower bound for A_k, we have the following result

Lemma 4 The sequence $\{A_k\}_{k\geq 0}$ defined by the recurrence (5.1) satisfies

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

Proof. We have:

$$\begin{aligned}\mu A_k A_{k+1} &\leq L(A_{k+1}-A_k)^2 = L(A_{k+1}^{1/2}-A_k^{1/2})^2(A_{k+1}^{1/2}+A_k^{1/2})^2 \\ &\leq 4LA_{k+1}(A_{k+1}^{1/2}-A_k^{1/2})^2.\end{aligned}$$

Therefore $\left(1+\frac{1}{2}\sqrt{\frac{\mu}{L}}\right)A_k^{1/2} \leq A_{k+1}^{1/2}$ which implies $A_k \geq \left(1+\frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}$. □

Concerning $\frac{\sum_{i=0}^k A_i}{A_k}$, the cumulative effect of the successive oracle errors, we begin with the following uniform upper-bound:

Lemma 5 The sequence $\{A_k\}_{k\geq 0}$ defined by the recurrence (5.1) satisfies

$$\frac{\sum_{i=0}^k A_i}{A_k} \leq 1+\frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}}+\sqrt{\frac{\mu}{L}+4}} \leq 1+\sqrt{\frac{L}{\mu}} \quad \forall k \geq 0.$$

Proof. We first note that A_k satisfies the following recurrence equation:

$$A_{k+1}^2 - \left(1+A_k\left(\frac{\mu}{L}\right)+2A_k\right)A_{k+1}+A_k^2 = 0$$

or equivalently:

$$A_{k+1} = \frac{\left(1+A_k\left(\frac{\mu}{L}\right)+2A_k\right)+\sqrt{\left(1+A_k\left(\frac{\mu}{L}\right)+2A_k\right)^2-4A_k^2}}{2}.$$

For our analysis, we consider also the sequence defined by the recurrence $\mu\tilde{A}_k = \frac{L(\tilde{A}_{k+1}-\tilde{A}_k)^2}{\tilde{A}_{k+1}}$ or equivalently $\tilde{A}_{k+1} = \tilde{A}_k\left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2}+1\right)$. We have clearly $\frac{A_{k+1}}{A_k} \geq \frac{\tilde{A}_{k+1}}{\tilde{A}_k} \quad \forall k \geq 0$ and therefore $\frac{A_k}{A_i} \geq \frac{\tilde{A}_k}{\tilde{A}_i} \quad \forall i<k, \quad \forall k \geq 1$. We conclude that:

$$\frac{\sum_{i=0}^k A_i}{A_k} \leq \frac{\sum_{i=0}^k \tilde{A}_i}{\tilde{A}_k}.$$

On the other hand, if we assume that $\tilde{A}_0 = A_0 = 1$, we have $\tilde{A}_k = C^k$ where $C = \left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2}+1\right)$. Therefore, we have $\sum_{i=0}^k \tilde{A}_i = \sum_{i=0}^k C^i = \frac{C^{k+1}-1}{C-1}$ and

$$\begin{aligned}\frac{\sum_{i=0}^k \tilde{A}_i}{\tilde{A}_k} &= \frac{C^{k+1}-1}{C-1}\frac{1}{C^k} = \frac{C^{k+1}-1}{C^{k+1}-C^k} \\ &\leq \frac{C}{C-1} = \frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}+2}{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}} = 1+\frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}}+\sqrt{\frac{\mu}{L}+4}} \\ &\leq 1+\sqrt{\frac{L}{\mu}}.\end{aligned}$$

We conclude that:

$$\frac{\sum_{i=0}^k A_i}{A_k} \leq \frac{\sum_{i=0}^k \tilde{A}_i}{\tilde{A}_k} \leq 1+\frac{2\sqrt{\frac{L}{\mu}}}{\sqrt{\frac{\mu}{L}}+\sqrt{\frac{\mu}{L}+4}} \leq 1+\sqrt{\frac{L}{\mu}}.$$

□

A pessimistic but also an optimistic interpretation of this result can be done:

- The FGM is worse than the GM concerning the effect of the oracle errors. Contrarily to the gradient method for which the oracle accuracy δ can be chosen of the same level that the desired final accuracy, the fast-gradient method suffers from an increase of error. The cumulative error (in the convergence rate for $f(y_k) - f^*$) coming from the successive oracle errors is bigger than each individual oracle error δ. This bad phenomenon does not come from our analysis but is a unavoidable problem of any fast first-order method for smooth strongly convex problems as we will see in Theorem 8. More precisely, for any optimal method in smooth strongly convex optimization, the total effect on the convergence rate of the successive oracle errors cannot be bounded by a uniform quantity (i.e. independent of k) having a better dependence in the condition number than $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$.
- When $\mu > 0$, the FGM does not suffer from an unbounded accumulation of errors. When $\mu = 0$, i.e. when the function is endowed with a (δ, L)-oracle, we have established in [1] that the fast gradient method suffers from an accumulation of oracle errors with rate $\Theta(k\delta)$, making the method asymptotically divergent. When $\mu > 0$, we can bound the total effect of the oracle errors by a quantity of order $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ that does not depend of k. This method is not divergent, the error on the function value converges to a limit smaller than $\left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$.

Of course, the cumulative effect of the oracle errors does not reach the level $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ from the very first iterations. In fact, we will prove now that the effect of the oracle errors is always less undesirable in the case $\mu > 0$ than in the case $\mu = 0$. We can bound $\frac{\sum_{i=0}^{k} A_i}{A_k}$ by an uniform quantity (impossible when $\mu = 0$) but also by the quantity of order $\Theta(k\Theta)$ available in the case $\mu = 0$:

Lemma 6 Let $\mu > 0$. The sequences $\{A_k(\mu)\}_{k\geq 0}$ and $\{A_k(0)\}_{k\geq 0}$ defined by the recurrences:

$$L + \mu A_k(\mu) = \frac{L(A_{k+1}(\mu) - A_k(\mu))^2}{A_{k+1}(\mu)}, \quad A_0(\mu) = 1$$

$$A_{k+1}(0) = (A_{k+1}(0) - A_k(0))^2, \quad A_0(0) = 1$$

satisfy:

$$\frac{\sum_{i=0}^{k} A_i(\mu)}{A_k(\mu)} \leq \frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)}.$$

In order to prove this result, we first establish the following lemma:

Lemma 7 For all $\mu > 0$, we have: $\frac{1}{A_k(\mu)} + \frac{\mu}{L} \geq \frac{1}{A_k(0)}$ i.e.: $A_k(0) \geq \frac{LA_k(\mu)}{L + A_k(\mu)\mu}$.

Proof. • It is true for $k = 0$. Indeed as $A_0(0) = A_0(\mu) = 1$, we have

$$\frac{1}{A_0(\mu)} + \frac{\mu}{L} \geq \frac{1}{A_0(0)}.$$

- Assume it is true for $k \geq 0$. We have:

$$\begin{aligned}
A_{k+1}(0) &= \frac{1+2A_k(0)+\sqrt{(1+2A_k(0))^2-4A_k(0)^2}}{2} \\
&= \frac{1+2A_k(0)+\sqrt{1+4A_k(0)}}{2} \\
&\geq \frac{1+\frac{2LA_k(\mu)}{L+\mu A_k(\mu)}+\sqrt{1+\frac{4LA_k(\mu)}{L+\mu A_k(\mu)}}}{2} \\
&= \frac{L+\mu A_k(\mu)+2LA_k(\mu)}{2(L+\mu A_k(\mu))} \\
&\quad +\frac{\sqrt{(L+A_k(\mu)\mu)^2+4LA_k(\mu)(L+A_k(\mu)\mu)}}{2(L+\mu A_k(\mu))} \\
&= \frac{L+(\mu)\,A_k(\mu)+2LA_k(\mu)}{2(L+\mu A_k(\mu))} \\
&\quad +\frac{\sqrt{(L+A_k(\mu)\mu+2L^2A_k(\mu))^2-4L^2A_k(\mu)^2}}{2(L+A_k(\mu)\mu)} \\
&= \frac{LA_{k+1}(\mu)}{L+\mu A_k(\mu)} \\
&\geq \frac{LA_{k+1}(\mu)}{L+\mu A_{k+1}(\mu)}.
\end{aligned}$$

since $A_{k+1}(\mu) \geq A_k(\mu)$.

□

We are now able to give the proof of the Lemma 6:

Proof. We have:

$$\frac{A_{k+1}(0)}{A_k(0)} = \frac{\frac{1}{A_k(0)}+2+\sqrt{\left(\frac{1}{A_k(0)}+2\right)^2-4}}{2}$$

and

$$\frac{A_{k+1}(\mu)}{A_k(\mu)} = \frac{\frac{1}{A_k(\mu)}+\frac{\mu}{L}+2+\sqrt{\left(\frac{1}{A_k(\mu)}+\frac{\mu}{L}+2\right)^2-4}}{2}.$$

Therefore as

$$\frac{1}{A_k(0)} \leq \frac{1}{A_k(\mu)}+\frac{\mu}{L}$$

using Lemma 7, we have:

$$\frac{A_{k+1}(0)}{A_k(0)} \leq \frac{A_{k+1}(\mu)}{A_k(\mu)}, \quad \forall k \geq 0.$$

As a consequence, we obtain:

$$\frac{A_k(0)}{A_i(0)} \leq \frac{A_k(\mu)}{A_i(\mu)}, \quad \forall 0 \geq i < k$$

and therefore

$$\frac{\sum_{i=0}^{k} A_i(\mu)}{A_k(\mu)} \leq \frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)}.$$

□

Remark 9 The behavior of the FGM in the case $\mu > 0$ is never worse than in the case $\mu = 0$. This is true for the rate of error accumulation, as we have seen in the Theorem 6, but also for the convergence rate in the exact case (i.e. the first term of the convergence rate, the term that does not depend on δ). Indeed, since $A_k(\mu) \geq A_k(0)$ for all $k \geq 0$, the first term in the convergence rate of the FGM i.e. $\frac{Ld(x^*)}{A_k(\mu)}$ can be bounded by $\frac{Ld(x^*)}{\left(1+\frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}}$ as we have seen in the Theorem 4 but also by the upper-bound $\frac{Ld(x^*)}{A_k(0)} = \Theta\left(\frac{LR^2}{k^2}\right)$ available in the case $\mu = 0$ (see [1]).

Remark 10 The fast gradient method presented here is compatible with the case $\mu = 0$ but is not completely equivalent in the choice of the sequence $\{\alpha_i\}_{i\geq 0}$ with the version analyzed in [1]. However, the two methods present the same rate of convergence and the same rate of errors accumulation. More precisely, in the FGM presented here, we have

$$\begin{aligned} A_{k+1}(0) &= (A_{k+1}(0) - A_k(0))^2 \\ &= (A_{k+1}(0)^{1/2} - A_k(0)^{1/2})^2 (A_{k+1}(0)^{1/2} + A_k(0)^{1/2})^2 \\ &\leq 4A_{k+1}(0)(A_{k+1}(0)^{1/2} - A_k(0)^{1/2})^2 \end{aligned}$$

and therefore:

$$A_{k+1}(0)^{1/2} \geq \frac{1}{2} + A_k(0)^{1/2}$$

which implies:

$$A_{k+1}(0) \geq \frac{(k+1)^2}{4}.$$

Furthermore, it is possible to show that

$$\frac{\sum_{i=0}^{k} A_i(0)}{A_k(0)} \leq \frac{1}{3}k + 2.4.$$

We conclude that

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{k^2} + (\frac{1}{3}k + 2.4)\delta$$

and we retrieve the classical behavior of a fast-gradient method when used with a $(\delta, L) = (\delta, L, 0)$-oracle (see [1]):

- A needed number of iterations in $\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$
- A needed oracle accuracy of order at least $\delta = \Theta\left(\frac{\epsilon^{3/2}}{L^{1/2}R}\right)$.

Now we can obtain the following convergence rate for the fast gradient method using a (δ, L, μ)-oracle

Theorem 7 The fast gradient method applied to a function f endowed with a (δ, L, μ)-oracle generates a sequence $\{y_k\}_{k\geq 1}$ satisfying:

$$\begin{aligned} f(y_k) - f^* &\leq \min\left(\frac{4Ld(x^*)}{k^2}, Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\quad + \min\left((\frac{1}{3}k + 2.4), \left(1 + \sqrt{\frac{L}{\mu}}\right)\right)\delta. \end{aligned}$$

Proof. Using the Lemmas 4 and 6 and the remarks 9 and 10 in the convergence rate given by the Theorem 6, we obtain

$$f(y_k) - f^* \leq \min\left(\frac{4Ld(x^*)}{k^2}, \frac{Ld(x^*)}{(1+\frac{1}{2}\sqrt{\frac{\mu}{L}})^{2k}}\right) + \min\left((\frac{1}{3}k+2.4), \left(1+\sqrt{\frac{L}{\mu}}\right)\right)\delta.$$

As for $x \in [0, \frac{1}{4}]$, we have that $\log(1+2x) \geq x$:

$$\begin{aligned}\frac{1}{(1+\frac{1}{2}\sqrt{\frac{\mu}{L}})^{2k}} &= \exp\left(-2k\log\left(1+\frac{1}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\leq \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right).\end{aligned}$$

We conclude that

$$\begin{aligned}f(y_k) - f^* &\leq \min\left(\frac{4Ld(x^*)}{k^2}, Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)\right) \\ &\quad + \min\left((\frac{1}{3}k+2.4), \left(1+\sqrt{\frac{L}{\mu}}\right)\right)\delta.\end{aligned}$$

□

5.3 Oracle accuracy fixed: Best reachable target accuracy using the FGM

We have seen that, contrarily to the GM, the FGM suffers from a problem of error increase. The extra error in the convergence rate, due to the (δ, L, μ) oracle, is not δ but something of order $\min\left(k\delta, \sqrt{\frac{L}{\mu}}\delta\right)$. As a consequence, the best possible level of accuracy δ cannot be reached by the FGM. In this section, we are interested in the best accuracy ϵ that we can obtain for $f(y_k) - f^*$, using the FGM with a (δ, L, μ) oracle.

We ensure here that δ, L, μ and R are fixed quantities. The only degree of freedom that we have is the number of iterations k that we perform. In Theorem 7, we have obtained the following model for the convergence rate of the FGM when applied to a function endowed with a (δ, L, μ) oracle:

$$f(y_k) - f^* \leq F(k) := \min\left(F_1(k), F_2(k), F_3(k), F_4(k)\right)$$

where

1. $F_1(k) = \frac{4Ld(x^*)}{k^2} + \left(\frac{1}{3}k+2.4\right)\delta$
2. $F_2(k) = \frac{4Ld(x^*)}{k^2} + \left(1+\sqrt{\frac{L}{\mu}}\right)\delta$
3. $F_3(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(\frac{1}{3}k+2.4\right)\delta$
4. $F_4(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(1+\sqrt{\frac{L}{\mu}}\right)\delta.$

The minimum of $F_1(k)$ is reached at $k_1^* = \Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$ and $F_1^* = F_1(k_1^*) = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$.

The minimum of $F_2(k)$, $F_2^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$, is reached at the limit. However we can obtain accuracy of the same order after $k_2^* = \Theta\left(\frac{(L\mu)^{1/4}R}{\delta^{1/2}}\right)$ iterations.

The minimum of $F_3(k)$ is reached at $k_3^* = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$ and $F_3^* = F_3(k_3^*) = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\left(\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)+1\right)\right)$.

The minimum of $F_4(k)$, $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ is reached at the limit. However we can obtain accuracy of the same order after $k_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$ iterations.

We conclude that the best reachable accuracy by the FGM is of order $\min\{F_1^*, F_4^*\} = \min\{\Theta(L^{1/3}R^{2/3}\delta^{2/3}), \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)\}$. More precisely, we can consider two different situations:

- A) The condition number $Q = \frac{L}{\mu}$ is sufficiently small and/or the oracle accuracy δ is sufficiently small and/or R is sufficiently big such that $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right) \leq F_1^* = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ i.e. $\delta \leq \Theta\left(\frac{R^2\mu^{3/2}}{L^{1/2}}\right)$.
  In this situation, we can gain from the fact that $\mu > 0$ and reach a level of accuracy $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ better than the best level reachable in the case $\mu = 0$ (i.e. $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$). The number of iterations needed in order to reach this accuracy is of order $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$.

  Remark 11 We also can reach the level $\Theta(L^{1/3}R^{2/3}\delta^{2/3})$ and the needed number of iterations is $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)\right)$. This number of iterations is smaller or of the same order as $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$, the needed number of iterations by the FGM to reach this accuracy in the case $\mu = 0$.

  Remark 12 We can also reach the level of accuracy $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ using the GM. But the needed number of iteration is of order $\Theta\left(\frac{L}{\mu}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$, worse than $\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{\sqrt{L\mu}R^2}{\delta}\right)\right)$.

- B) The condition number $Q = \frac{L}{\mu}$ is sufficiently big and/or the oracle accuracy δ is sufficiently big and/or R is sufficiently small such that $F_4^* = \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right) \geq F_1^* = \Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ i.e. $\delta \geq \Theta\left(\frac{R^2\mu^{3/2}}{L^{1/2}}\right)$.
  In this situation, we cannot exploit the fact that $\mu > 0$. The best reachable accuracy is of level $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ after $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$ iterations. It is the same result as what we had obtained in [1] in the case $\mu = 0$.

  Remark 13 We can reach also reach the level $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$. The needed number of iteration is $\Theta\left(\frac{(L\mu)^{1/4}R}{\delta^{1/2}}\right)$.

  Remark 14 We can also reach the level of accuracy $F_1^* = \Theta(L^{1/3}R^{2/3}\delta^{2/3})$ using the GM. But the needed number of iterations is of order $\min\{\Theta\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)\right)\} = \Theta\left(\frac{L^{2/3}R^{4/3}}{\delta^{2/3}}\right)$ which is worse than $\Theta\left(\frac{L^{1/3}R^{2/3}}{\delta^{1/3}}\right)$.

Remark 15 The impossibility to exploit the fact that $\mu > 0$ when μ is too small comes perhaps from our analysis. More precisely, it might come from the fact that we have bounded $f(y_k) - f^* \leq \frac{1}{A_k}\left(d(x^*) + \sum_{i=0}^k A_i\delta\right)$ using the two approximations:

1. $1 + \mu A_k \approx 1$ when μ is small
2. $1 + \mu A_k \approx \mu A_k$ when μ is big.

It would seems more natural for the best reachable accuracy, to be a continuous function with limits $\Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)$ when $\mu \to +\infty$ and $\Theta\left(L^{1/3}R^{2/3}\delta^{2/3}\right)$ when $\mu \to 0$. However, it seems very difficult to find an upper bound for $f(y_k) - f^*$ which is at the same time accurate and easy to analyze.

In conclusion, the best accuracy reachable by the FGM when endowed with a (δ, L, μ) oracle is of order $\min\{\Theta(L^{1/3}R^{2/3}\delta^{2/3}), \Theta\left(\sqrt{\frac{L}{\mu}}\delta\right)\}$. If such accuracy is sufficient, it is preferable to use the FGM instead of the GM. However, if we want to reach a better level of accuracy, for example of order $\Theta(\delta)$, the only possibility is to use the GM, slower but less sensitive to the oracle error.

5.4 Oracle accuracy not fixed: Required number of iterations and required oracle accuracy for a given target accuracy

In this subsection, we consider a different situation. We assume that we can choose the number of iterations k and the oracle accuracy δ but that we want to reach a target accuracy ϵ for the objective function. Furthermore, in this case we assume that L and μ are independent of the oracle accuracy δ. A way to ensure $f(y_k) - f^* \leq \epsilon$ is to choose k and δ such that one of the four models $F_i(k)$ is smaller than ϵ. First, we will consider the four models separately. Each model contains three terms and we will ensure $F_i(k) \leq \epsilon$ by imposing that each term in the model is smaller than $\frac{\epsilon}{3}$ (another repartition of the desired accuracy between the different terms of a model leads simply to different constant factors in the resulting expressions for k and δ).

1. Model 1: $F_1(k) = \frac{4Ld(x^*)}{k^2} + \left(\frac{1}{3}k + 2.4\right)\delta$.
   We have the three conditions ensuring $F_1(k) \leq \epsilon$:
   - $$\frac{4Ld(x^*)}{k^2} \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}.$$
     We choose $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$
   - $$\frac{1}{3}k\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}}$$
   - $$2.4\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon}{7.2}.$$

   Therefore a first possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$ iterations with $\delta = \min\{\frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}}, \frac{\epsilon}{7.2}\}$.
2. Model 2: $F_2(k) = \frac{4Ld(x^*)}{k^2} + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$.
   We have the three conditions ensuring $F_2(k) \leq \epsilon$:
   - $$\frac{4Ld(x^*)}{k^2} \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
   - $$\delta \leq \frac{\epsilon}{3}$$

- $$\sqrt{\frac{L}{\mu}}\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon.$$

Therefore a second possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$ iterations with $\delta = \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon$.

3. Model 3: $F_3(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(\frac{1}{3}k + 2.4\right)\delta$.
We have the three conditions ensuring $F_3(k) \leq \epsilon$:

- $$Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right).$$

We choose $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$.

- $$\frac{1}{3}k\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$

- $$2.4\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{\epsilon}{7.2}.$$

Therefore a third condition ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$ iterations with $\delta = \min\{\frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}, \frac{\epsilon}{7.2}\}$.

4. Model 4: $F_4(k) = Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$. We have the three conditions ensuring $F_4(k) \leq \epsilon$:

- $$Ld(x^*)\exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) \leq \frac{\epsilon}{3} \Leftrightarrow k \geq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$$

- $$\delta \leq \frac{\epsilon}{3}$$

- $$\sqrt{\frac{L}{\mu}}\delta \leq \frac{\epsilon}{3} \Leftrightarrow \delta \leq \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon.$$

Therefore a fourth possibility for ensuring $f(y_k) - f^* \leq \epsilon$ is to perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$ with $\delta = \frac{1}{3}\sqrt{\frac{\mu}{L}}\epsilon$.

Of course, we want to reach $f(y_k) - f^* \leq \epsilon$ in a minimum number of iterations k and with δ as big as possible (δ representing the accuracy of the first-order information, it seems natural that a high accuracy for δ is costly). We will choose between these four possibilities that ensure $f(y_k) - f^* \leq \epsilon$ with the minimization of k as a first criterion and the maximization of δ as a second criterion.

Remark 16 For simplicity, we assume here that $\epsilon \leq 0.2315Ld(x^*)$ i.e. $\frac{\epsilon^{3/2}}{2\sqrt{3}L^{1/2}d(x^*)^{1/2}} \leq \frac{\epsilon}{7.2}$. In particular, this assumption implies:

$$\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \geq \frac{3}{2}.$$

We consider two main cases :

1. **Case 1:**
$$2\sqrt{\frac{3Ld(x^*)}{\epsilon}} \leq 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right)$$
i.e.
$$\sqrt{\frac{L}{\mu}} \geq \frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$
In this case, models 1 and 2 are the most favorable regarding to the number of iterations. We perform $k = 2\sqrt{\frac{3Ld(x^*)}{\epsilon}} = \Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right)$ iterations. Concerning the needed oracle accuracy, we have to consider two different subcases:
   - **Case 1.1:**
$$\frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)} \leq \sqrt{\frac{L}{\mu}} \leq \frac{2}{3}\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
In this case $\sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} \geq \frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}}$ and model 2 is more interesting than the first one. We choose $\delta = \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$.
   - **Case 1.2:**
$$\sqrt{\frac{L}{\mu}} \geq \frac{2}{3}\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
In this case $\frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}} \geq \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3}$ and model 1 is more interesting than the second one. We choose $\delta = \frac{\epsilon^{3/2}}{L^{1/2}\sqrt{d(x^*)}2\sqrt{3}} = \Theta\left(\frac{\epsilon^{3/2}}{L^{1/2}R}\right)$.

   Remark 17 : In the case 1.2., we do not exploit the fact that $\mu > 0$. We obtain the same number of iterations and the same oracle accuracy in the case $\mu = 0$.

2. **Case 2:**
$$2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \leq 2\sqrt{\frac{3Ld(x^*)}{\epsilon}}$$
i.e.:
$$\sqrt{\frac{L}{\mu}} \leq \frac{\sqrt{\frac{3Ld(x^*)}{\epsilon}}}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)}.$$
In this case, models 3 and 4 are the most favorable with respect to the needed number of iterations. We perform $k = 2\sqrt{\frac{L}{\mu}}\log\left(\frac{3Ld(x^*)}{\epsilon}\right) = \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{3LR^2}{\epsilon}\right)\right)$ iterations. As $\log\left(\frac{3Ld(x^*)}{\epsilon}\right) \geq \frac{3}{2}$, we have
$$\frac{1}{2}\sqrt{\frac{\mu}{L}}\frac{\epsilon}{\log\left(\frac{3Ld(x^*)}{\epsilon}\right)} \leq \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3}.$$
Therefore model 4 is always more interesting than the third one and we choose $\delta = \sqrt{\frac{\mu}{L}}\frac{\epsilon}{3} = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$.

6 GM and FGM for uniformly convex problems with different levels of smoothness

In the Sections 3 and 5, we have studied the effect of a (δ, L, μ) oracle on the GM and the FGM. In particular, we have established the complexity of these methods in a inexact framework and the link between desired final accuracy and needed oracle accuracy.

In subsection 2.5, we have seen that a (δ, L, μ) oracle can be also available for functions that are not in $S^{1,1}_{\mu,L}(Q)$. More precisely, the function can be uniformly convex (instead of strongly convex) and nonsmooth or weakly smooth (instead of smooth). The exact oracle for a function $f \in U^{0,\rho,\nu}_{\kappa,M}(Q)$ can be seen as a (δ, L, μ) oracle.

If we put these results together, we can apply the GM and the FGM, initially designed for functions in $S^{1,1}_{\mu,L}(Q)$, to functions in $U^{0,\rho,\nu}_{\kappa,M}(Q)$. In this section, we study the complexity of these two methods on various classes of convex problems with different levels of smoothness and different levels of uniform convexity. For simplicity, we are only interested in the order of the dependence of these complexities on ϵ (the desired final accuracy), κ and M, not on the absolute constant factors. Furthermore, Theorem 3 is applied with $\delta_1 = \delta_2 = \frac{\delta}{2}$.

6.1 Gradient method for function in $U^{0,\rho,\nu}_{\kappa,M}(Q)$.

If we apply the gradient method to a function endowed with a (δ, L, μ) oracle and if the desired accuracy is $\epsilon > 0$, we know that the number of iterations that we have to perform is

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

with an oracle accuracy $\delta = \Theta(\epsilon)$. When the (δ, L, μ) oracle is in fact an exact oracle of a function $f \in U^{0,\rho,\mu}_{\kappa,M}(Q)$, we have (see Theorem 3) $L = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\delta^{\frac{1-\nu}{1+\nu}}}\right)$ and $\mu = \Theta\left(\kappa^{\frac{2}{\rho}}\delta^{\frac{\rho-2}{\rho}}\right)$. Therefore

$$\frac{L}{\mu} = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\delta^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\right)$$

and

$$\log\left(\frac{LR^2}{\epsilon}\right) = \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

We obtain the complexity:

$$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

We can particularize this complexity bound for different classes of uniformly convex problems:

1. The smooth case $\nu = 1$ (f has a Lipschitz-continuous gradient)
   - Strong convexity $\rho = 2$:
   $$\Theta\left(\frac{M}{\kappa}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$
   We retrieve the non-optimal complexity of the gradient method on $S^{1,1}_{\kappa,M}(Q)$.

- Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{M}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{\rho-2}{\rho}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$

This complexity cannot be optimal (see what we obtain in the next subsection with the FGM).

2. The nonsmooth case $\nu = 0$ (f has subgradients with bounded variation)
   - Strong convexity $\rho = 2$:

$$\Theta\left(\frac{M^2}{\kappa\epsilon}\log\left(\frac{M^2R^2}{\epsilon^2}\right)\right)$$

This complexity is optimal up to a logarithmic factor (see [3, 2]).

   - Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{M^2}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{2(\rho-1)}{\rho}}}\log\left(\frac{M^2R^2}{\epsilon^2}\right)\right)$$

This complexity is optimal, up to a logarithmic factor (see [2]).

3. The weakly smooth case $0 < \nu < 1$ (f has a Hölder continuous gradient)
   - Strong convexity $\rho = 2$:

$$\Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa\epsilon^{\frac{1-\nu}{1+\nu}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right)$$

This complexity cannot be optimal (see what we obtain with FGM in the next subsection).

   - Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right) = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\kappa^{\frac{2}{\rho}}\epsilon^{\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}}}\log\left(\frac{M^{\frac{2}{1+\nu}}R^2}{\epsilon^{\frac{2}{1+\nu}}}\right)\right).$$

6.2 Fast gradient method for functions in $U^{0,\rho,\nu}_{\kappa,M}(Q)$

If we apply the fast gradient method to a function endowed with a (δ, L, μ) oracle and if the desired accuracy is ϵ, we know that the number of iterations that we have to perform is proportional to

$$\Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)$$

with an oracle accuracy $\delta = \Theta\left(\sqrt{\frac{\mu}{L}}\epsilon\right)$. When the (δ, L, μ) oracle is in fact an exact oracle of a function $f \in U^{0,\rho,\mu}_{\kappa,M}(Q)$, we have (see Theorem 3) $L = \Theta\left(\frac{M^{\frac{2}{1+\nu}}}{\delta^{\frac{1-\nu}{1+\nu}}}\right)$ and $\mu = \Theta\left(\kappa^{\frac{2}{\rho}}\delta^{\frac{\rho-2}{\rho}}\right)$. We obtain:

$$\sqrt{\frac{L}{\mu}} = \Theta\left(\frac{M^{\frac{1}{1+\nu}}}{\kappa^{\frac{1}{\rho}}\delta^{\frac{1}{2}\left(\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}\right)}}\right)$$

and therefore

$$\sqrt{\frac{L}{\mu}}\delta = \Theta\left(\frac{M^{\frac{1}{1+\nu}}\delta^{1-\frac{1}{2}\left(\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}\right)}}{\kappa^{\frac{1}{\rho}}}\right).$$

As $\sqrt{\frac{L}{\mu}}\delta = \Theta(\epsilon)$ and as $1-\frac{1}{2}\left(\frac{1-\nu}{1+\nu}+\frac{\rho-2}{\rho}\right) = \frac{\nu}{\nu+1}+\frac{1}{\rho}$, we obtain that $\delta = \Theta\left(\left(\frac{\kappa^{\frac{1}{\rho}}\epsilon}{M^{\frac{1}{1+\nu}}}\right)^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\right)$. Therefore, we have

$$\sqrt{\frac{L}{\mu}} = \Theta\left(\frac{\epsilon}{\delta}\right) = \Theta\left(\frac{M^{\frac{1}{1+\nu}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}}{\kappa^{\frac{1}{\rho}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\epsilon^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}-1}}\right)$$

and

$$\begin{aligned}\log\left(\frac{LR^2}{\epsilon}\right) &= \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}}}{\left(\left(\frac{\kappa^{\frac{1}{\rho}}\epsilon}{M^{\frac{1}{1+\nu}}}\right)^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\right)^{\frac{1-\nu}{1+\nu}}}\frac{R^2}{\epsilon}\right)\right)\\ &= \Theta\left(\log\left(\frac{M^{\frac{2}{1+\nu}+\frac{1-\nu}{(1+\nu)^2}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}R^2}{\kappa^{\frac{1}{\rho}\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}\epsilon^{\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)+1}}\right)\right).\end{aligned}$$

We obtain the complexity:

$$\Theta\left(\frac{M^{\frac{1}{1+\nu}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}}{\kappa^{\frac{1}{\rho}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\epsilon^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}-1}}\log\left(\frac{M^{\frac{2}{1+\nu}+\frac{1-\nu}{(1+\nu)^2}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}R^2}{\kappa^{\frac{1}{\rho}\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}\epsilon^{\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)+1}}\right)\right).$$

We particularize now this complexity bound on different classes of uniformly convex problems:

1. The smooth case $\nu = 1$ (f has a Lipschitz-continuous gradient)
   - Strong convexity $\rho = 2$:
   $$\Theta\left(\frac{M^{\frac{1}{2}}}{\kappa^{\frac{1}{2}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$
   We retrieve the optimal complexity of the fast gradient method on $S^{1,1}_{\kappa,M}(Q)$.
   - Uniform convexity $\rho \geq 2$:
   $$\Theta\left(\frac{M^{\frac{\rho}{\rho+2}}}{\kappa^{\frac{2}{\rho+2}}\epsilon^{\frac{\rho-2}{\rho+2}}}\log\left(\frac{MR^2}{\epsilon}\right)\right)$$
2. The nonsmooth case $\nu = 0$ (f has subgradients with bounded variation)
   - Strong convexity $\rho = 2$:
   $$\Theta\left(\frac{M^2}{\kappa\epsilon}\log\left(\frac{M^4R^2}{\kappa\epsilon^3}\right)\right)$$
   This complexity is optimal up to a logarithmic factor (see [3, 2]).
   - Uniform convexity $\rho \geq 2$:
   $$\Theta\left(\frac{M^\rho}{\kappa\epsilon^{\rho-1}}\log\left(\frac{M^{2+\rho}R^2}{\kappa\epsilon^{\rho+1}}\right)\right)$$
   This complexity is clearly non-optimal (compare with what we obtain using the GM in the previous subsection).

3. The weakly smooth case $0 < \nu < 1$ (f has a Hölder continuous gradient)
   - Strong convexity $\rho = 2$:

$$\Theta\left(\frac{M^{\frac{2}{1+3\nu}}}{\kappa^{\frac{\nu+1}{3\nu+1}}\epsilon^{\frac{1-\nu}{1+3\nu}}}\log\left(\frac{M^{\frac{4\nu+4}{(1+\nu)(1+3\nu)}}R^2}{\kappa^{\frac{1-\nu}{3\nu+1}}\epsilon^{\frac{3+\nu}{1+3\nu}}}\right)\right)$$

   - Uniform convexity $\rho \geq 2$:

$$\Theta\left(\frac{M^{\frac{1}{1+\nu}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}}{\kappa^{\frac{1}{\rho}\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}}\epsilon^{\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}-1}}\log\left(\frac{M^{\frac{2}{1+\nu}+\frac{1-\nu}{(1+\nu)^2}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}R^2}{\kappa^{\frac{1}{\rho}\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)}\epsilon^{\frac{1-\nu}{1+\nu}\left(\frac{1}{\frac{\nu}{\nu+1}+\frac{1}{\rho}}\right)+1}}\right)\right).$$

7 Lower bound on errors increase

Applicability of first-order methods, initially designed for smooth strongly convex problems, to nonsmooth strongly convex problems, using the notion of inexact oracle, opens a possibility to derive lower bounds on error increase. This is the main subject of this section.

Theorem 8 Consider a first-order method for $S^{1,1}_{\mu,L}(Q)$. Assume that the bounds on the performance of this method, as applied to a problem equipped with a (δ, L, μ)-oracle, are given by inequality

$$f(z_k) - f^* \leq \min\left(C_1\frac{LR^2}{k^{p_1}}, C_2LR^2\exp\left(-k\left(\frac{\mu}{L}\right)^{p_2}\right)\right) + \min\left(C_3k^{q_1}\delta, C_4\left(\frac{L}{\mu}\right)^{q_2}\delta\right). \tag{7.1}$$

where C_1, C_2, C_3, C_4 are absolute constants, and k is the iteration counter. Then the inequalities

$$q_1 \geq p_1 - 1$$

and

$$q_2 \geq 1 - p_2$$

must hold.

Proof.
- $q_1 \geq p_1 - 1$.

  Let f be a nonsmooth convex function, whose subgradients have variation bounded by constant M i.e $f \in F^{0,0}_M(Q)$. We have seen in [1] that for such a function, the standard oracle can be treated as a $(\delta, \frac{M^2}{2\delta}, 0)$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$f(z_k) - f^* \quad \leq \quad \frac{C_1M^2R^2}{2\delta k_1^p} + C_3k^{q_1}\delta.$$

  Optimizing the right-hand side of this inequality in δ, we get

$$f(z_k) - f^* \quad \leq \quad [2C_1C_3]^{1/2}MR\cdot k^{-\frac{p_1-q_1}{2}}.$$

  From the lower complexity bounds for nonsmooth optimization problems, we know that black-box methods cannot converge faster than $O(\frac{1}{k^{1/2}})$. Hence, we conclude that $p_1 - q_1 \leq 1$.

- $q_2 \geq 1 - p_2$.

  Let f be a nonsmooth strongly convex function, whose subgradients have variation bounded by constant M i.e. $f \in U^{1,2,0}_{\mu,M}(Q)$. We have seen in Theorem 3 that for such

a function, the standard oracle can be treated as $(\delta, \frac{M^2}{2\delta}, \mu)$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$\begin{aligned} f(z_k) - f^* &\leq \frac{C_2 M^2 R^2}{2\delta} \exp\left(-k\left(\frac{\mu}{M^2} 2\delta\right)^{p_2}\right) + C_4 \left(\frac{M^2}{2\delta\mu}\right)^{q_2} \delta \\ &= \frac{C_2 M^2 R^2}{2\delta} \exp\left(-k\frac{\mu^{p_2} 2^{p_2} \delta^{p_2}}{M^{2p_2}}\right) + C_4 \frac{M^{2q_2}}{2^{q_2}\mu^q} \delta^{1-q_2}. \end{aligned}$$

If we choose δ such that $C_4 \frac{M^{2q_2}}{2^{q_2}\mu^{q_2}} \delta^{1-q_2} = \frac{\epsilon}{2}$, we obtain $\delta(\epsilon) = \frac{1}{2} \frac{\mu^{\frac{q_2}{1-q_2}} \epsilon^{\frac{1}{1-q_2}}}{C_4^{\frac{1}{1-q_2}} M^{\frac{2q_2}{1-q_2}}}$. Therefore, if we want an accuracy of ϵ for the objective function, we can choose k such that $\frac{C_2 M^2 R^2}{2\delta(\epsilon)} \exp\left(-k\frac{\mu^{p_2} 2^{p_2} \delta^{p_2}}{M^{2p_2}}\right) = \frac{\epsilon}{2}$ i.e.

$$k = \frac{M^{\frac{2p_2}{1-q_2}} C_4^{\frac{p_2}{1-q_2}}}{\mu^{\frac{p_2}{1-q_2}} \epsilon^{\frac{p_2}{1-q_2}}} \log\left(\frac{2C_2 M^{\frac{2}{1-q_2}} R^2}{\epsilon^{\frac{2-q_2}{1-q_2}} \mu^{\frac{q_2}{1-q_2}}}\right).$$

From the lower complexity bounds for nonsmooth strongly convex optimization problems, we know that black-box methods cannot have a better complexity that $O\left(\frac{1}{\epsilon}\right)$ (see [3, 2]). Hence, we conclude that $\frac{p_2}{1-q_2} \geq 1 \quad \Leftrightarrow p_2 \geq 1 - q_2$.

□

We can consider two extreme cases:

- $q_1 = 0$ and $q_2 = 0 \Rightarrow p_1 \leq 1$ and $p_2 \geq 1$:
  It is impossible to have a first-order method without increase of errors, which has better complexity than GM, that is
  $\min\left(\Theta\left(\frac{LR^2}{\epsilon}\right), \Theta\left(\frac{L}{\mu}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$.
- $p_1 = 2$ and $p_2 = \frac{1}{2} \Rightarrow q_1 \geq 1$ and $q_2 \geq \frac{1}{2}$:
  If we want a first-order method with optimal complexity
  $\min\left(\Theta\left(\sqrt{\frac{LR^2}{\epsilon}}\right), \Theta\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{LR^2}{\epsilon}\right)\right)\right)$ like the FGM, then it must suffer from increase of errors, with factor at least of order $\min\left(\Theta(k), \Theta\left(\sqrt{\frac{L}{\mu}}\right)\right)$ (we obtain exactly this factor for the FGM).

References

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

[2] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *Technical Report*, (2010).

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

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

[5] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Serie A*, **103(1)**, 127-152 (2005).

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