Noname manuscript No.
(will be inserted by the editor)

First-Order Methods of Smooth Convex Optimization with Inexact Oracle.

Olivier Devolder · François Glineur · Yurii Nesterov

the date of receipt and acceptance should be inserted later

Abstract We introduce the notion of inexact first-order oracle and analyze the behaviour of several first-order methods of smooth convex optimization used with such an oracle. This notion of inexact oracle naturally appears in the context of smoothing techniques, Moreau-Yosida regularization, Augmented Lagrangians and many other situations.

We derive complexity estimates for primal, dual and fast gradient methods, and study in particular their dependence on the accuracy of the oracle and the desired accuracy of the objective function. We observe that the superiority of fast gradient methods over the classical ones is no longer absolute when an inexact oracle is used. We prove that, contrary to simple gradient schemes, fast gradient methods must necessarily suffer from error accumulation.

Finally, we show that the notion of inexact oracle allows the application of first-order methods of smooth convex optimization to solve non-smooth or weakly smooth convex problems.

The first author is a F.R.S.-FNRS Research Fellow. 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 scientific responsibility rests with its authors.

O. Devolder
Université catholique de Louvain, ICTEAM Institute, B-1348 Louvain-La-Neuve, Belgium;
Université catholique de Louvain, CORE, B-1348 Louvain-La-Neuve, Belgium.
Tel.: +32-10-479422, Fax: +32-10-474301
E-mail: Olivier.Devolder@uclouvain.be

F. Glineur
Université catholique de Louvain, ICTEAM Institute, B-1348 Louvain-La-Neuve, Belgium;
Université catholique de Louvain, CORE, B-1348 Louvain-La-Neuve, Belgium.
E-mail: Francois.Glineur@uclouvain.be

Y. Nesterov
Université catholique de Louvain, ICTEAM Institute, B-1348 Louvain-La-Neuve, Belgium;
Université catholique de Louvain, CORE, B-1348 Louvain-La-Neuve, Belgium.
E-mail: Yurii.Nesterov@uclouvain.be

Keywords Smooth convex optimization, first-order methods, inexact oracle, gradient methods, fast gradient methods, complexity bounds.

Mathematics Subject Classification (2000) 90C06, 90C25, 90C60

1 Introduction

In large-scale convex optimization, first-order methods are methods of choice due to their cheap iteration cost. When the objective function is assumed to be smooth, for example when its gradient is Lipschitz-continuous with constant L, the simplest numerical schemes to be considered are the gradient method and its variants. If accuracy ϵ is desired for the objective function, these methods require $O\left(\frac{L}{\epsilon}\right)$ iterations.

However, it is well-known that in the black-box framework [11], first-order methods can achieve the lower complexity bound of $O\left(\sqrt{\frac{L}{\epsilon}}\right)$ iterations. Such optimal methods, called Fast Gradient Methods (FGM), have been developed for various classes of problems since 1983 [12,13,14,15] and outperform theoretically, and often in practice, the classical gradient methods. Interest into these methods has been renewed recently with development of smoothing techniques for non-smooth convex problems (see [15,16,17,4]), where FGMs are used to minimize a smooth approximation of a non-smooth objective function.

Standard analysis of first-order methods assumes availability of exact first-order information. Namely, the oracle must provide at each given point the exact values of the function and its gradient. However, in many convex problems, including those obtained by smoothing techniques, the objective function and its gradient are computed by solving another auxiliary optimization problem. In practice, we are often only able to solve these subproblems approximately. Hence, in that context, numerical methods solving the outer problem are provided with inexact first-order information. This led us to investigate the behavior of first-order methods working with an inexact oracle.

We introduce in Section 2 a new definition of inexact first-order oracle and list a few simple examples. In Section 3, we show how our concept is applicable to situations when the inexact oracle is computed by an auxiliary optimization problem. In particular, we consider convex-concave saddle point problems, augmented Lagrangians, and Moreau-Yosida regularization.

In Sections 4 and 5, we consider classical (primal and dual) and fast gradient methods, designed for the class of convex functions with Lipschitz-continuous gradient. We obtain efficiency estimates when these methods are used with an inexact first-order oracle. We also study the link between desired accuracy for the objective function and necessary accuracy for the oracle. We observe that the superiority of the fast gradient methods over the classical ones is no longer absolute when an inexact oracle is used, because FGMs suffer from error accumulation. In particular, fast methods require first-order information with higher accuracy than standard gradient methods to obtain a solution with a given accuracy. Therefore, the choice between these methods depends

on the availability and relative cost of an inexact oracle at different levels of accuracy, as is explained in Section 6.

In Section 7, we compare our approach with other definitions of inexact oracle, as applied to the smoothed max-representable functions typically obtained by the smoothing techniques [3, 1]. We show that our definition can give better complexity results.

Our definition of inexact oracle is applicable to non-smooth and weakly smooth convex problems. Section 8 shows how to apply first-order methods designed for smooth convex optimization to functions with a weaker level of smoothness. For that, we show that (exact) first-order information for a non-smooth problem, such as subgradients, can be viewed as an inexact oracle, so that the methods of Sections 4 and 5 can be applied. We obtain in this way "universal" first-order methods possessing optimal rates of convergence for objective functions with different level of smoothness. We also prove lower bounds on the rate of error accumulation for any first-order method using an inexact oracle, which shows that all methods discussed in this paper have the lowest possible rate of error accumulation. In particular, it appears that while slower standard gradient methods are able to maintain an error comparable to the oracle accuracy, any optimal method must suffer from error accumulation.

2 Inexact first-order oracle

2.1 Motivation and definition

We consider the following convex optimization problem:

$$f^* = \min_{x \in Q} f(x), \tag{1}$$

where Q is a closed convex set in a finite-dimensional space E, and function f is convex on Q. Space E is endowed with the norm $\|\cdot\|_E$ and E^*, the dual space of E, with the dual norm $\|g\|_E^* = \sup_{y \in E}\{|\langle g, y\rangle| : \|y\|_E \le 1\}$ where $\langle ., .\rangle$ denotes the dual pairing. For example, by fixing a positive definite self-adjoint operator $B : E \to E^*$, we can define the following Euclidean norms:

$$\|h\|_E = \|h\|_2 = \langle Bh, h\rangle \quad \forall h \in E$$

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

We assume that problem (1) is solvable with optimal solution x^*.

Consider $F_L^{1,1}(Q)$, the class of convex functions on convex set Q whose gradient is Lipschitz-continuous with constant L. It is well-known that functions belonging to this class satisfy

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

see the top of Figure 1. Moreover, it is easy to check that, for a given y, quantities $f(y)$ and $\nabla f(y)$ are uniquely determined by this pair of inequalities.

Therefore, membership in $F_L^{1,1}(Q)$ can be characterized by the existence of an *oracle* returning for each point $y \in Q$ a pair $(f_L(y), g_L(y)) \in \mathbb{R}\times E^*$, necessarily equal to $(f(y), \nabla f(y))$, satisfying

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

(both zeroth-order and first-order information are included in the oracle). Our definition of an inexact oracle simply consists in introducing a given amount δ of tolerance in this pair of inequalities (see bottom of Figure 1).

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

$$0 \;\leq\; f(x) - \big(f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle\big) \leq\; \frac{L}{2}\|x-y\|_E^2 + \delta \text{ for all } x \in Q. \quad (3)$$

A function f belongs to $F_L^{1,1}(Q)$ if and only it admits a $(0, L)$-oracle, namely $(f_{0,L}(y), g_{0,L}(y)) = (f(y), \nabla f(y))$. However, the class of functions admitting a (δ, L)-oracle is strictly larger, and includes non-smooth functions, as we will see later.

2.2 Properties

We list here a few important properties of (δ, L)-oracles.

- A (δ, L)-oracle provides a lower δ-approximation of the function value. Indeed, taking $x = y$ in (3), we obtain

$$f_{\delta,L}(y) \leq f(y) \leq f_{\delta,L}(y) + \delta. \quad (4)$$

- A (δ, L)-oracle provides a δ-subgradient of f at $y \in Q$, i.e.

$$g_{\delta,L}(y) \in \partial_\delta f(y) = \{z \in E^* : f(x) \geq f(y) + \langle z, x-y\rangle - \delta \quad \forall x \in Q\}.$$

  Indeed, using the first inequality in (3) and (4), we have for all $x, y \in Q$

$$f(x) \;\geq\; f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle \;\geq\; f(y) + \langle g_{\delta,L}(y), x-y\rangle - \delta. \quad (5)$$

  Methods of non-smooth convex optimization based on δ-subgradients have a long history (see e.g. [20,19,2,9] for subgradient methods, and [2,6,7] for proximal point and bundle methods). We will show later that a standard subgradient can also satisfy the second inequality in (3), which opens the possibility of using the concept of inexact oracle in the context of non-smooth convex optimization.
- A (δ, L) oracle can certify than an approximate solution has accuracy δ. Indeed, assuming $g_{\delta,L}(y)$ satisfies $\langle g_{\delta,L}(y), x-y\rangle \geq 0$ for all $x \in Q$, we have that $f_{\delta,L}(y) \leq f(x^*) = f^*$ and therefore, using (4), we have $f(y) \leq f^* + \delta$.

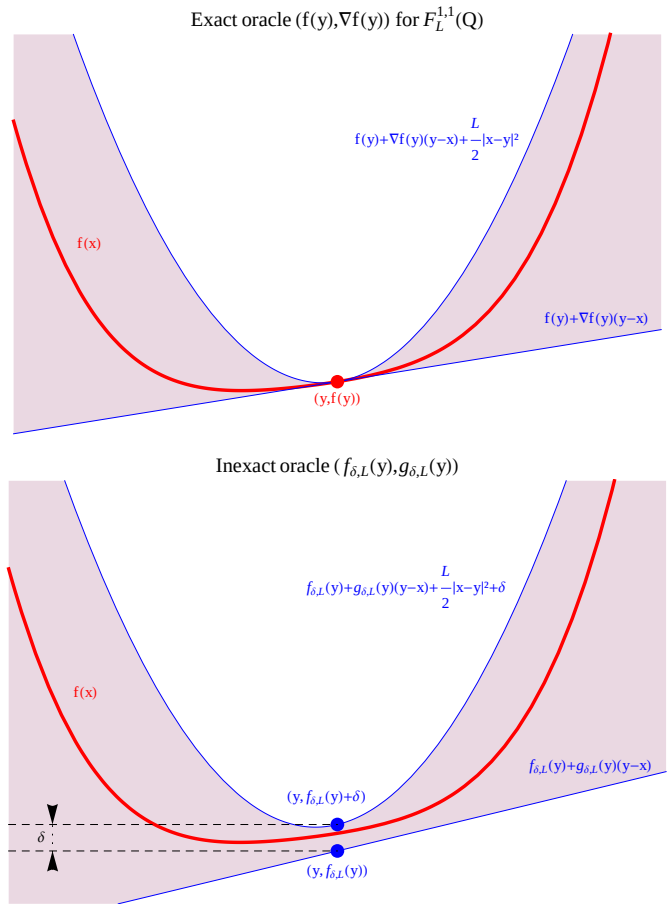

Fig. 1 Illustration of lower and upper bounds (blue lines) implied by the definition of an exact (top) and inexact (bottom) oracle.

– If f admits a (δ, L)-oracle, then cf admits a $(c\delta, cL)$-oracle for any value of the constant $c > 0$. If f_i admits a (δ_i, L_i)-oracle, $i = 1, 2$, then $f_1 + f_2$ admits a $(\delta_1 + \delta_2, L_1 + L_2)$-oracle.
– When $Q = E$, the difference between $g_{\delta,L}$ and any subgradient $g_y \in \partial f(y)$ is bounded as follows

$$\|g_y - g_{\delta,L}(y)\|_E^* \leq [2\delta L]^{\frac{1}{2}}. \tag{6}$$

Indeed, for any $x \in Q$ we have $f(x) \geq f(y) + \langle g_y, x - y\rangle \geq f_{\delta,L}(y) + \langle g_y, x - y\rangle$. Subtracting this inequality from the second part of (3), we get that

$$\langle g_y - g_{\delta,L}(y), x - y\rangle \leq \frac{L}{2}\|x - y\|_E^2 + \delta$$

holds for all $x \in Q$. If $z \in E$ is such that $\|g_y - g_{\delta,L}(y)\|_E^* = |\langle g_y - g_{\delta,L}(y), z\rangle|$ and $\|z\|_E = 1$ and if we choose $x \in Q$ such that $x - y = t\ \mathrm{sign}(\langle g_y - g_{\delta,L}, z\rangle)z$

with $t > 0$, we obtain:

$$t\|g_y - g_{\delta,L}(y)\|_E^* \leq \frac{L}{2}t^2 + \delta \Leftrightarrow \|g_y - g_{\delta,L}(y)\|_E^* \leq \frac{L}{2}t + \frac{\delta}{t}\ . \qquad (7)$$

This upper bound attains its minimum $[2\delta L]^{\frac{1}{2}}$ when $t = [\frac{2\delta}{L}]^{\frac{1}{2}}$. In particular, when $Q = E$, parameter t is free to take any real value, and we obtain inequality (6). For constrained problems, a similar bound can be obtained in terms of the distance $d(y, \partial Q)$ between y and the boundary of Q: letting

$$d(y, \partial Q) = \max\{r | \|x - y\|_E \leq r \Rightarrow x \in Q\}$$

we have that (7) holds for all t such that $0 < t \leq d(y, \partial Q)$, so that

$$\|g_y - g_{\delta,L}(y)\|_E^* \leq \begin{cases} \frac{L}{2}d(y,\partial Q) + \frac{\delta}{d(y,\partial Q)} & \text{when } 0 < d(y,\partial Q) \leq [\frac{2\delta}{L}]^{\frac{1}{2}}\ , \\ [2\delta L]^{\frac{1}{2}} & \text{when } d(y,\partial Q) \geq [\frac{2\delta}{L}]^{\frac{1}{2}}\ . \end{cases}$$

– If E is endowed with the Euclidean norm $\|.\|_2$, the distance between exact and inexact gradient mappings can be bounded by the same quantities as the distance between exact and inexact (sub)gradients. Recall that for any $\gamma > 0$, $g \in E^*$ and $y \in E$, the gradient mapping $M_\gamma(y, g)$, which replaces the gradient for constrained problems, is defined by

$$T_\gamma(y, g) = \arg\min_{x \in Q}\{\langle g, x - y\rangle + \frac{\gamma}{2}\|x - y\|_E^2\} \qquad (8)$$

$$M_\gamma(y, g) = \gamma(y - T_\gamma(y, g)). \qquad (9)$$

If f is subdifferentiable at point y, the exact gradient mapping for any subgradient $g_y \in \partial f(y)$ is equal to $M_\gamma(y, g_y)$. Similarly, if an inexact (δ, L) oracle returns $(f_{\delta,L}(y), g_{\delta,L}(y))$ for point y, we call $M_\gamma(y, g_{\delta,L}(y))$ the inexact gradient mapping. We are going to prove that the following holds

$$\|M_\gamma(y, g_y) - M_\gamma(y, g_{\delta,L}(y))\|_2 \leq \|g_y - g_{\delta,L}(y)\|_2^*\,. \qquad (10)$$

First-order optimality conditions for (8) can be written as

$$\langle g + \gamma B(T_\gamma(y, g) - y), x - T_\gamma(y, g)\rangle \geq 0 \quad \forall x \in Q. \qquad (11)$$

Applying those to $T_\gamma(y, g_y)$ and $T_\gamma(y, g_{\delta,L}(y))$ leads to

$$\begin{aligned} \langle g_y - BM_\gamma(y, g_y), x - T_\gamma(y, g_y)\rangle &\geq 0 \quad \forall x \in Q \\ \langle g_{\delta,L}(y) - BM_\gamma(y, g_{\delta,L}(y)), x - T_\gamma(y, g_{\delta,L}(y))\rangle &\geq 0 \quad \forall x \in Q \end{aligned}$$

and specializing respectively to $x = T_\gamma(y, g_{\delta,L}(y))$ and $x = T_\gamma(y, g_y)$ gives

$$\begin{aligned} \langle g_y - BM_\gamma(y, g_y), T_\gamma(y, g_{\delta,L}(y)) - T_\gamma(y, g_y)\rangle &\geq 0 \\ \langle g_{\delta,L}(y) - BM_\gamma(y, g_{\delta,L}(y)), T_\gamma(y, g_y) - T_\gamma(y, g_{\delta,L}(y))\rangle &\geq 0\ . \end{aligned}$$

Using now (9) in the inner products, multiplying by γ and summing, we obtain

$$\langle g_y - BM_\gamma(y, g_y) - g_{\delta,L}(y) + BM_\gamma(y, g_{\delta,L}(y)), M_\gamma(y, g_y) - M_\gamma(y, g_{\delta,L}(y))\rangle \geq 0$$

and assuming that E is endowed with the Euclidean norm $\|.\|_2$ gives

$$\langle g_y - g_{\delta,L}(y), M_\gamma(y,g_y) - M_\gamma(y,g_{\delta,L}(y))\rangle \geq \|M_\gamma(y,g_y) - M_\gamma(y,g_{\delta,L}(y))\|_2^2 ,$$

from which the desired inequality (10) follows by Cauchy-Schwartz.

Characterizing the class of functions that can be endowed with a (δ, L)-oracle is an interesting open question. We provide below some necessary conditions in the simple case where $Q = E$ and E is endowed with the Euclidean norm $\|.\|_2$. First of all, we establish the following inequality:

Theorem 1 *If f is equipped with a (δ, L)-oracle, we have*

$$\frac{1}{2L}(\|g_{\delta,L}(x) - g_{\delta,L}(y)\|_2^*)^2 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta \quad \forall x, y \in E.$$

Proof Let $x \in E$ and consider the function $F(y) = f(y) - \langle g_{\delta,L}(x), y\rangle$.
As $(f_{\delta,L}(y), g_{\delta,L}(y))$ is a (δ, L)-oracle for f and $(-\langle g_{\delta,L}(x), y\rangle, -g_{\delta,L}(x))$ is a $(0,0)$-oracle for $-\langle g_{\delta,L}(x), y\rangle$, the resulting sum of oracles $(F_{\delta,L}(y), G_{\delta,L}(y)) = (f_{\delta,L}(y) - \langle g_{\delta,L}(x), y\rangle, g_{\delta,L}(y) - g_{\delta,L}(x))$ is a (δ, L)-oracle for $F(y)$. Using the lower bound in the definition of the oracle $F_{\delta,L}(x) + \langle G_{\delta,L}(x), y - x\rangle \leq F(y)$, valid for any y, and the fact that $G_{\delta,L}(x) = 0$, we derive

$$\begin{aligned} F_{\delta,L}(x) &\leq F\left(y - \frac{1}{L}B^{-1}G_{\delta,L}(y)\right) \\ &\leq F_{\delta,L}(y) + \langle G_{\delta,L}(y), -\frac{1}{L}B^{-1}G_{\delta,L}(y)\rangle + \frac{L}{2}\left(\left\|\frac{1}{L}G_{\delta,L}(y)\right\|_2^*\right)^2 + \delta \\ &= F_{\delta,L}(y) - \frac{1}{2L}\left(\|G_{\delta,L}(y)\|_2^*\right)^2 + \delta \end{aligned}$$

which allows us to obtain

$$\begin{aligned} \frac{1}{2L}\left(\|g_{\delta,L}(y) - g_{\delta,L}(x)\|_2^*\right)^2 &\leq f_{\delta,L}(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta \\ &\leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle + \delta. \end{aligned}$$

□

As a Corollary, we have:

Corollary 1 *If f is equipped with a (δ, L)-oracle, then we have for all $x, y \in E$*

$$\|g_{\delta,L}(x) - g_{\delta,L}(y)\|_2^* \leq \sqrt{L^2\|x - y\|_2^2 + 4L\delta}$$

and for any $g_x \in \partial f(x)$ and any $g_y \in \partial f(y)$

$$\|g_x - g_y\|_2^* \leq (2\sqrt{2} + 2)\sqrt{L\delta} + L\|x - y\|_2 .$$

Proof Our first claim directly follows from

$$\begin{aligned}\frac{1}{2L}\left(\|g_{\delta,L}(x)-g_{\delta,L}(y)\|_2^*\right)^2 &\leq f(y)-f_{\delta,L}(x)-\langle g_{\delta,L}(x),y-x\rangle+\delta\\ &\overset{(3)}{\leq} \frac{L}{2}\|x-y\|_2^2+2\delta.\end{aligned}$$

Furthermore, for any $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$, we have by (6):

$$\|g_x - g_{\delta,L}(x)\|_2^* \leq \sqrt{2\delta L},$$

$$\|g_y - g_{\delta,L}(y)\|_2^* \leq \sqrt{2\delta L},$$

and therefore (using the $\|\cdot\|_2 \leq \|\cdot\|_1$ inequality for the last step)

$$\begin{aligned}\|g_x-g_y\|_2^* &\leq \|g_x-g_{\delta,L}(x)\|_2^*+\|g_{\delta,L}(x)-g_{\delta,L}(x)\|_2^*+\|g_{\delta,L}(y)-g_y\|_2^*\\ &\leq 2\sqrt{2\delta L}+\sqrt{L^2\|x-y\|_2^2+4L\delta}\\ &\leq (2\sqrt{2}+2)\sqrt{\delta L}+L\|x-y\|_2.\end{aligned}$$

□

We conclude that the variation of subgradients of f is locally bounded, i.e.

$$\|g_x-g_y\|_2^* \leq (2\sqrt{2}+2)\sqrt{L\delta}+LR \quad \forall\, x,y \text{ s.t. } \|x-y\|_2 \leq R.$$

Note however that this property is true for any subdifferentiable convex function defined on the whole space E. Assume now that function f is endowed with a family of $(\delta, L(\delta))$-oracles and consider the following situations:

– $$\lim_{\delta\to 0} L(\delta) = \bar{L} < +\infty$$

In this case we have $\|g_x - g_y\|_2^* \leq \bar{L}\|x-y\|_2$ and f must be a smooth convex function with a Lipschitz-continuous gradient.

– $$\lim_{\delta\to\infty} L(\delta) = 0 \text{ and } \lim_{\delta\to\infty} L(\delta)\delta = \bar{C} < +\infty,$$

which is the case for example when $L(\delta) = \frac{\bar{C}}{\delta}$. We have $\|g_x - g_y\|_2^* \leq (2\sqrt{2}+2)\sqrt{\bar{C}}$ so that f must be a convex function with bounded variation of subgradients.

– $$\lim_{\delta\to\infty} L(\delta) = 0 \text{ and } \lim_{\delta\to\infty} L(\delta)\delta = 0$$

which would happen for example when $L(\delta) = \frac{\bar{C}}{\delta^2}$. We have in that case that $\|g_x - g_y\|_2^* \leq 0$ and f must be a constant function.

2.3 Examples

To conlude this section, we consider four simple examples of inexact oracle. More sophisticated examples will be given in Section 3.

a. Computations at shifted points. Let function $f \in F_M^{1,1}(Q)$ be endowed with an oracle providing at each point $y \in Q$ the exact values of function and gradient, albeit computed at a shifted point $\hat{y}$ different from y. Let us show that such an oracle can be converted into a (δ, L)-oracle with

$$\delta = M \|y - \hat{y}\|_E^2, \quad L = 2M.$$

Convexity of f implies the following inequality for any $x \in Q$

$$\begin{aligned} f(x) &\geq f(\hat{y}) + \langle \nabla f(\hat{y}), x - \hat{y} \rangle \\ &= f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \langle \nabla f(\hat{y}), x - y \rangle. \end{aligned}$$

Therefore, to satisfy the first inequality in (3) we can choose $f_{\delta,L}(y) \stackrel{\text{def}}{=} f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle$, and $g_{\delta,L}(y) \stackrel{\text{def}}{=} \nabla f(\hat{y})$. In order to prove the second inequality in (3), note that we have for all $x \in Q$

$$\begin{aligned} f(x) &\stackrel{(2)}{\leq} f(\hat{y}) + \langle \nabla f(\hat{y}), x - \hat{y} \rangle + \tfrac{M}{2} \|x - \hat{y}\|_E^2 \\ &= f(\hat{y}) + \langle \nabla f(\hat{y}), y - \hat{y} \rangle + \langle \nabla f(\hat{y}), x - y \rangle + \tfrac{M}{2} \|x - \hat{y}\|_E^2. \end{aligned}$$

Since $\|\cdot\|_E^2$ is a convex function, we have

$$\begin{aligned} \|x - \hat{y}\|_E^2 &= \left\| \frac{1}{2}\big(2(x - y)\big) + \frac{1}{2}\big(2(y - \hat{y})\big) \right\|_E^2 \qquad (12) \\ &\leq \frac{1}{2} \|2(x - y)\|_E^2 + \frac{1}{2} \|2(y - \hat{y})\|_E^2 = 2\|y - \hat{y}\|_E^2 + 2\|x - y\|_E^2 \qquad (13) \end{aligned}$$

Therefore,

$$f(x) \leq f_{\delta,L}(y) + \langle g_{\delta,L}(y), x - y \rangle + M \|x - y\|_E^2 + M \|y - \hat{y}\|_E^2.$$

We can therefore choose $L = 2M$ and $\delta = M\|y - \hat{y}\|_E^2$ to satisfy the (δ, L)-oracle definition.

b. Convex problems with weaker level of smoothness. Let us show that the notion of (δ, L)-oracle can be useful for solving the problems with *exact* first-order information but with a lower level of smoothness. Let function f be convex and subdifferentiable on Q. For each $y \in Q$, denote by $g(y)$ an arbitrary element of the subdifferential $\partial f(y)$. Assume that f satisfies the following Hölder condition:

$$\|g(x) - g(y)\|_E^* \leq L_\nu \|x - y\|_E^\nu, \quad \forall x, y \in Q, \tag{14}$$

where $\nu \in [0, 1]$, and $L_\nu < +\infty$. This condition leads to the following inequality:

$$f(x) \leq f(y) + \langle g(y), x - y\rangle + \tfrac{L_\nu}{1+\nu} \|x - y\|_E^{1+\nu}, \quad \forall x, y \in Q. \tag{15}$$

Denote the class of such functions by $F_{L_\nu}^{1,\nu}(Q)$. When $\nu = 1$, we get functions with Lipschitz-continuous gradient. For $\nu < 1$, we get a lower level of smoothness. In particular, when $\nu = 0$, we obtain functions whose subgradients have *bounded variation*. Clearly, the latter class includes functions whose subgradients are uniformly bounded by M (just take $L_0 = 2M$).

Let us fix $\nu \in [0, 1)$ and an arbitrary $\delta > 0$. We are going to find a constant $A(\delta, \nu)$ such that for any function $f \in F_{L_\nu}^{1,\nu}(Q)$ we have

$$f(x) - f(y) - \langle g(y), x - y\rangle \leq \tfrac{A(\delta,\nu)}{2} \|x - y\|_E^2 + \delta, \quad \forall x, y \in Q. \tag{16}$$

Comparing (15) and (16), we need choose $A(\delta, \nu)$ such that

$$\frac{L_\nu}{1+\nu} \|x - y\|_E^{1+\nu} \;\leq\; \frac{A(\delta,\nu)}{2} \|x - y\|_E^2 + \delta \,.$$

Since $t = \|x - y\|_E^2$ can take any nonnegative value, we may choose

$$A(\delta, \nu) = 2 \max_{t \geq 0} \left\{ \tfrac{L_\nu}{1+\nu} t^{-1+\nu} - \delta t^{-2} \right\} = L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}}.$$

(the latter expression is obtained after straightforward computations, the optimal value of t in the maximization being $t_* = \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\nu} \right]^{-\frac{1}{1+\nu}}$). This means that the exact first-order information $(f(y), g(y))$ also constitutes an inexact $(\delta, A(\delta, \nu))$-oracle. We will therefore be able to apply the methods from Sections 4 and 5, initially devoted for smooth problems, to the minimization of the non- or weakly smooth objective f.

For example, for functions with bounded variation of subgradients ($\nu = 0$) we have

$$A(\delta, 0) = \tfrac{L_0^2}{2\delta}. \tag{17}$$

so that a $(\delta, \frac{L_0^2}{2\delta})$-oracle is available for all values of $\delta > 0$.

Note that parameter δ does not represent an actual accuracy: it can be chosen arbitrarily, *independently* of the answer of the oracle. In particular, δ can be chosen as small as we want, at the price of a larger value for Lipschitz constant L of the (δ, L)-oracle, which grows as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$. Section 8 will describe the details and consequences of the application of first-order method of smooth convex optimization to non-smooth or weakly smooth functions.

Remark 1 This analysis can easily be extended to the case where δ-subgradients with bounded variations are used instead of exact subgradients. We obtain in this case a $(2\delta, A(\delta,\nu))$-oracle.

c. Function smoothed by local averaging. Another typical approach in order to apply first-order method of $F_L^{1,1}(E)$ to a non-smooth function consists in smoothing the function by averaging of first-order information. Assume that E is endowed with an Euclidean norm and consider a non-smooth convex function $f \in F_M^{1,0}(E)$. Let $r > 0, y \in E$, and define

$$f_\delta(y) = \frac{1}{V_r} \int_{\|z-y\|_2 \le r} f(z)\,\mathrm{d}z$$
$$\nabla f_r(y) = g_r(y) = \frac{1}{V_r} \int_{\|z-y\|_2 \le r} g(z)\,\mathrm{d}z$$

where V_r denotes the volume of a Euclidean ball with radius r, and $\{g(z) : \|z-y\|_2 \le r\}$ is a measurable selection of subgradients of f in this ball. As f is convex and Lipschitz-continuous with constant M we have

$$0 \le f(x) - f(z) - \langle g(z), x-z\rangle \le M\|x-z\|_2 \quad \forall x,z \in E$$

and therefore

$$f(x) \ge f(z) + \langle g(z), x-y\rangle + \langle g(z), y-z\rangle \quad \forall x,y,z \in E$$
$$f(x) \le f(z) + \langle g(z), x-y\rangle + \langle g(z), y-z\rangle + M\|x-z\|_2 \quad \forall x,y,z \in E.$$

Averaging now these two inequalities with respect to z over the ball $\{z : \|z-y\|_2 \le r\}$, we obtain for all $x, y \in Z$

$$f(x) \ge f_r(y) + \langle g_r(y), x-y\rangle - Mr$$
$$f(x) \le f_r(y) + \langle g_r(y), x-y\rangle + Mr + \frac{M}{V_r}\int_{\|z-y\|_2 \le r} \|x-z\|_2\,\mathrm{d}z$$

(where we used that $|\langle g(z), y-z\rangle| \le \|g(z)\|_2^* \|y-z\|_2 \le Mr$). Furthermore, we have

$$\|x-z\|_2 \overset{(13)}{\le} \sqrt{2\|x-y\|_2^2 + 2\|z-y\|_2^2} \le \frac{2\|x-y\|_2^2 + 2\|z-y\|_2^2}{2r} + \frac{r}{2}$$

(where the second inequality comes from the arithmetic-geometric inequality), and therefore

$$f(x) \le f_r(y) + \langle g_r(y), x-y\rangle + \frac{3}{2}Mr + M\frac{\|x-y\|_2^2}{r} + \frac{M}{V_r}\int_{\|z-y\|_2 \le r} \|z-y\|_2\,\mathrm{d}z$$
$$\le f_r(y) + \langle g_r(y), x-y\rangle + \frac{5}{2}Mr + M\frac{\|x-y\|_2^2}{r}.$$

Finally, choosing $f_{\delta,L}(y) = f_r(y) - Mr$, $g_{\delta,L}(y) = g_r(y)$, $\delta = \frac{7Mr}{2}$ and $L = \frac{2M}{r}$, we obtain a $(\delta, L) = (\frac{7Mr}{2}, \frac{2M}{r}) = (\delta, \frac{7M^2}{\delta})$-oracle. Note that the dependence of L in M and δ is similar to that of the previous example, where subgradients are used directly instead of being averaged.

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

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

We can then show that

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

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

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

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

One might wonder whether all inexact oracles can be obtained in that fashion, i.e. whether any inexact oracle can be seen as an exact oracle for a smooth approximation $\bar{f}$. It turns out that is not the case: indeed, as we have seen earlier, when f has subgradients with bounded variation, its exact function values and subgradients can be seen as a (δ, L)-oracle (for arbitrary value of δ). Clearly, such an oracle cannot be at the same time equal to the exact function values and gradient of any smooth function $\bar{f}$.

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

3 Inexact oracle for functions defined by an optimization problem

3.1 Accuracy measures for approximate solutions

In this section, we consider smooth convex optimization problems of the form (1) whose objective function is defined by another optimization problem:

$$f(x) = \max_{u \in U} \Psi(x, u), \tag{18}$$

where U is a convex set of a finite dimensional space F endowed with the norm $\|.\|_F$ and for any $x \in Q$ function $\Psi(x, \cdot)$ is smooth and (strongly) concave with concavity parameter $\kappa \geq 0$. Computation of f and its gradient requires the exact solution of this auxiliary problem. However, in practice, such a solution might often be impossible or too costly to compute, so that an approximate solution has to be used instead.

We will measure the accuracy of an approximate solution u_x for problem (18) in three different ways:

$$V_1(u_x) = \max_{u\in U}\langle \nabla_2\Psi(x,u_x), u-u_x\rangle,$$

$$V_2(u_x) = \max_{u\in U}\left[\Psi(x,u)-\Psi(x,u_x)+\tfrac{\kappa}{2}\,\|u_x-u\|_F^2\right], \tag{19}$$

$$V_3(u_x) = \max_{u\in U}\left[\Psi(x,u)-\Psi(x,u_x)\right].$$

Since $\Psi(x,\cdot)$ is strongly concave, we have:

$$\Psi(x,u) \le \Psi(x,u_x) + \langle \nabla_2\Psi(x,u_x), u-u_x\rangle - \tfrac{\kappa}{2}\|u-u_x\|_F^2, \quad \forall\, u\in U.$$

Therefore our three measures are related by

$$V_3(u_x) \le V_2(u_x) \;\le\; V_1(u_x).$$

For a given level of accuracy $\delta>0$, the condition $V_1(u_x)\le\delta$ is the strongest, and condition $V_3(u_x)\le\delta$ is the most relaxed.

We describe below three classes of max-type functions for which the approximate solution of subproblem (18), when satisfying one of the conditions $V_i(u_x)\le\delta$, allows the construction of a (δ,L)-oracle.

Let us show first how to satisfy stopping criteria (19) in practice. The most common criterion is the third one. It amounts to estimating the optimality gap in the value of objective function. Many optimization methods offer direct control of this criterion. Other criteria might be more difficult to handle. Therefore, let us describe a "brute force" approach designed to satisfy the strongest V_1 criterion (Here we assume that F is endowed with an Euclidean norm).

Let $D_u<\infty$ be the diameter of U. Let us choose $u_0\in U$ and form a new function

$$\bar\Psi(x,u) = \Psi(x,u) - \tfrac{1}{2}\mu\|u-u_0\|_2^2.$$

Denote by $\bar V_i(u)$ the corresponding accuracy measures, and $u_x^* = \arg\max\limits_{u\in U}\bar\Psi(x,u)$. For any $u\in U$ we obtain

$$\begin{aligned}
0 \ge \langle \nabla_2\bar\Psi(x,u_x^*), u-u_x^*\rangle \;&=\; \langle \nabla_2\bar\Psi(x,u_x^*), u_x-u_x^*\rangle + \langle \nabla_2\bar\Psi(x,u_x^*), u-u_x\rangle \\
&\ge -\bar V_3(u_x) + \langle \nabla_2\bar\Psi(x,u_x^*) - \nabla_2\bar\Psi(x,u_x), u-u_x\rangle + \langle \nabla_2\bar\Psi(x,u_x), u-u_x\rangle \\
&\ge -\bar V_3(u_x) - \|\nabla_2\bar\Psi(x,u_x^*) - \nabla_2\bar\Psi(x,u_x)\|_2^* D_u + \langle \nabla_2\bar\Psi(x,u_x), u-u_x\rangle.
\end{aligned}$$

Since $\nabla_2\bar\Psi(x,\cdot)$ is Lipschitz continuous on U with constant L, we get

$$\begin{aligned}
\frac{1}{2L}\left(\left\|\nabla_2\bar\Psi(x,u_x^*) - \nabla_2\bar\Psi(x,u_x)\right\|_2^*\right)^2 &\le \bar\Psi(x,u_x^*) - \bar\Psi(x,u_x) + \langle \nabla_2\bar\Psi(x,u_x^*), u_x-u_x^*\rangle \\
&\le \bar\Psi(x,u_x^*) - \bar\Psi(x,u_x) = \bar V_3(u_x).
\end{aligned}$$

and therefore:

$$V_1(u_x) \;\leq\; \bar{V}_1(u_x) + \mu D_u^2 \overset{(2)}{\leq} \bar{V}_3(u_x) + D_u[2L\bar{V}_3(u_x)]^{1/2} + \mu D_u^2.$$

Thus, if we choose $\mu = \frac{\delta}{3D_u^2}$, we can get the desired level of $V_1(u_x)$ by ensuring $\bar{V}_3(u_x) \leq \frac{\delta^2}{18LD_u^2}$. Note that function $\bar{\Psi}(x,\cdot)$ is strongly concave. Therefore, the complexity of its maximization in the scale $\bar{V}_3$ depends logarithmically on the desired accuracy. If this is done, for example, by a FGM, it requires at most $O(\frac{L^{1/2}}{\delta^{1/2}} \ln \frac{1}{\delta})$ iterations (see section 2.2 in [14]).

3.2 Functions obtained by smoothing techniques

Let U be a closed, convex set of a finite dimensional space F endowed with the norm $\|\cdot\|_F$, and

$$\Psi(x,u) = G(u) + \langle Au, x\rangle,$$

where $A : F \to E^*$ is a linear operator, and $G(u)$ is a differentiable, strongly concave function with concavity parameter $\kappa > 0$. Under these assumptions, optimization problem (18) has only one optimal solution u_x^*. Moreover, f is convex and smooth with Lipschitz-continuous gradient $\nabla f(x) = Au_x^*$. The corresponding Lipschitz-constant is equal to

$$L(f) = \tfrac{1}{\kappa}\|A\|^2_{F\to E^*} \tag{20}$$

where $\|A\|_{F\to E^*} = \max\{\|Au\|_{E^*} : \|u\|_F = 1\}$. The importance of this class of functions is justified by the smoothing approach for non-smooth convex optimization (see [15,16,17,4]).

Suppose that for all $y \in Q$ we can find a point $u_y \in U$ satisfying condition

$$V_3(u_y) = \Psi(y, u_y^*) - \Psi(y, u_y) \;\leq\; \tfrac{\delta}{2}. \tag{21}$$

Let us show that this allows us to construct an $(\delta, 2L(f))$-oracle. Indeed, since $\Psi(\cdot, u)$ is convex, for all $u \in U$, we have

$$\begin{aligned} f(x) = \Psi(x,u_x^*) \;&\geq\; \Psi(x,u_y) \;\geq\; \Psi(y,u_y) + \langle \nabla_1\Psi(y,u_y), x-y\rangle \\ &= f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle, \end{aligned} \tag{22}$$

where $f_{\delta,L}(y) \overset{\text{def}}{=} \Psi(y,u_y)$, $g_{\delta,L}(y) \overset{\text{def}}{=} \nabla_1\Psi(y,u_y) = Au_y$, and L will be specified later. Further, note that

$$\langle \nabla_1\Psi(y,u_y^*), x-y\rangle = \langle g_{\delta,L}(y), x-y\rangle + \langle A(u_y^*-u_y), x-y\rangle. \tag{23}$$

Since f has Lipschitz-continuous gradient, we have:

$$\begin{aligned} f(x) \;&\leq\; f(y) + \langle \nabla f(y), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 \\ &=\; f(y) + \langle \nabla\Psi_1(y,u_y^*), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 \\ &\overset{(23)}{=}\; f(y) + \langle g_{\delta,L}(y), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 + \langle A(u_y^*-u_y), x-y\rangle. \end{aligned}$$

On the other hand, we have:

$$\begin{aligned}\langle A(u_y^* - u_y), x - y\rangle &\leq \left\|u_y^* - u_y\right\|_F \left\|A^T(x-y)\right\|_E^* \\ &\overset{(20)}{\leq} \tfrac{\kappa}{2}\left\|u_y^* - u_y\right\|_F^2 + \tfrac{L(f)}{2}\left\|x-y\right\|_E^2.\end{aligned}$$

Therefore,

$$f(x) \leq f(y) + \langle g_{\delta,L}(y), x - y\rangle + L(f)\left\|x-y\right\|_E^2 + \tfrac{\kappa}{2}\left\|u_y^* - u_y\right\|_F^2.$$

Since Ψ is strongly concave, $\frac{\kappa}{2}\left\|u_y - u_y^*\right\|_F^2 \leq \Psi(y, u_y^*) - \Psi(y, u_y)$. Thus,

$$f(x) \leq \Psi(y, u_y) + 2(\Psi(y, u_y^*) - \Psi(y, u_y)) + \langle g_{\delta,L}(y), x - y\rangle + L(f)\left\|x-y\right\|_E^2.$$

In view of conditions (21) and (22), we have proved that the pair $(\Psi(y, u_y), Au_y)$, satisfying condition (21), corresponds to an (δ, L)-oracle with $L = 2L(f)$.

3.3 Moreau-Yosida regularization

In this section, we consider functions of the form

$$f(x) = \min_{u\in U}\left\{\mathcal{L}(x,u) \overset{\text{def}}{=} h(u) + \tfrac{\kappa}{2}\left\|u - x\right\|_2^2\right\}, \tag{24}$$

where h is a smooth convex function on a convex set $U \subset \mathbb{R}^n$ endowed with the usual Euclidean norm $\left\|x\right\|_2^2 = \langle x, x\rangle$. The function f is convex with Lipschitz-continuous gradient $\nabla f(x) = \kappa(x - u_x^*)$, where u_x^* denotes the unique optimal solution of the problem (24). The Lipschitz constant of the gradient is equal to κ.

Instead of solving exactly the problem (24), we compute a feasible solution u_x satisfying

$$V_2(u_x) = \max_{u\in U}\left\{\mathcal{L}(x,u_x) - \mathcal{L}(x,u) + \tfrac{\kappa}{2}\left\|u - u_x\right\|_2^2\right\} \leq \delta. \tag{25}$$

(Since $\mathcal{L}$ is convex in u, we inverted the sign in the definition of V_2 in (19).) Let us show that for all $x \in Q$ the objects

$$\begin{aligned} f_{\delta,L}(x) &= \mathcal{L}(x,u_x) - \delta = h(u_x) + \tfrac{\kappa}{2}\left\|u_x - x\right\|_2^2 - \delta, \\ g_{\delta,L}(x) &= \nabla_1\mathcal{L}(x,u_x) = \kappa(x - u_x)\end{aligned} \tag{26}$$

correspond to an answer of an (δ, L)-oracle with $L = \kappa$. Indeed,

$$\begin{aligned} f(x) &= \mathcal{L}(x, u_x^*) \geq \mathcal{L}(y, u_x^*) + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - x - y\rangle \\ &\overset{(25)}{\geq} \mathcal{L}(y, u_y) + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - x - y\rangle \\ &= \mathcal{L}(y, u_y) + \kappa\langle y - u_y, x - y\rangle + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta \\ &\quad + \tfrac{\kappa}{2}\langle y - x, 2u_x^* - 2u_y + y - x\rangle \\ &= \mathcal{L}(y, u_y) + \kappa\langle y - u_y, x - y\rangle - \delta \\ &\quad + \tfrac{\kappa}{2}\left(\|u_x^* - u_y\|_2^2 + \|y - x\|_2^2 + 2\langle y - x, u_x^* - u_y\rangle\right) \\ &\geq \mathcal{L}(y, u_y) + \kappa\langle y - u_y, x - y\rangle - \delta. \end{aligned}$$

Thus, we satisfy the first inequality in (3) with the values defined by (26). Further, for all $x, y \in Q$ we have

$$\begin{aligned} f(x) = h(u_x^*) + \tfrac{\kappa}{2}\|u_x^* - x\|_2^2 &\leq h(u_y) + \tfrac{\kappa}{2}\|u_y - x\|_2^2 \\ &= h(u_y) + \tfrac{\kappa}{2}\|u_y - y\|_2^2 + \tfrac{\kappa}{2}\langle x - y, x + y - 2u_y\rangle \\ &= \mathcal{L}(y, u_y) + \kappa\langle y - u_y, x - y\rangle + \tfrac{\kappa}{2}\|y - x\|_2^2. \end{aligned}$$

Thus, in view of definition (26), we have proved the second inequality in (3) with $L = \kappa$.

3.4 Functions defined by Augmented Lagrangians

Consider the following convex problem:

$$\max_{u \in U} \{h(u) : \ Au = 0\}, \tag{27}$$

where h is a smooth concave function on the convex set $U \subset F$, F is a finite-dimensional space, and $A : F \to E^*$ is a linear operator. Let E be endowed with the Euclidean norm $\|.\|_2$. In the Augmented Lagrangian approach, we need to solve the dual problem

$$\min_{x \in E} \ f(x), \tag{28}$$

where

$$f(x) \overset{\text{def}}{=} \max_{u \in U} \left[\Psi(x, u) \overset{\text{def}}{=} h(u) + \langle Au, x\rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\right]. \tag{29}$$

It is well-known that f is a convex smooth function with Lipschitz-continuous gradient

$$\nabla f(x) = Au_x^*,$$

where u_x^* denotes any optimal solution of the optimization problem (29). The Lipschitz constant of the gradient is equal to $\frac{1}{\kappa}$.

Assume that, instead of solving (28) exactly, we compute an approximate solution $u_x \in U$ such that

$$\begin{aligned} V_1(u_x) &= \max_{u\in U} \langle \nabla_2 \Psi(x,u_x), u-u_x\rangle \\ &= \max_{u\in U} \langle \nabla h(u_x) + A^T x - \kappa A^T B^{-1} A u_x, u-u_x\rangle \ \le\ \delta. \end{aligned} \tag{30}$$

Let us show that the objects

$$f_{\delta,L}(x) \ =\ \Psi(x,u_x), \quad g_{\delta,L}(x) \ =\ \nabla_1 \Psi(x,u_x) \ =\ Au_x \tag{31}$$

correspond to a (δ, L)-oracle with $L = \frac{1}{\kappa}$. Indeed, for all $x, y \in E$ we have

$$\begin{aligned} f(x) &= \max_{u\in U} \left\{ h(u) + \langle Au, x\rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2 \right\} \\ &\ge h(u_y) + \langle Au_y, x\rangle - \tfrac{\kappa}{2}(\|Au_y\|_2^*)^2 \ =\ \Psi(y,u_y) + \langle Au_y, x-y\rangle. \end{aligned}$$

Thus, in view of definition (31), the first inequality in (3) is proved. Further,

$$\begin{aligned} f(x) \ &\le\ \max_{u\in U}\{h(u_y) + \langle \nabla h(u_y), u-u_y\rangle + \langle Au, x\rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\} \\ &\overset{(30)}{\le}\ \max_{u\in U}\{h(u_y) - \langle A^T y - \kappa A^T B^{-1} A u_y, u-u_y\rangle + \langle Au, x\rangle - \tfrac{\kappa}{2}(\|Au\|_2^*)^2\} + \delta \\ &=\ \Psi(y,u_y) + \langle Au_y, x-y\rangle \\ &\quad + \max_{u\in U}\left\{ \langle A(u-u_y), x-y\rangle - \tfrac{\kappa}{2}(\|A(u-u_y)\|_2^*)^2 \right\} + \delta. \end{aligned}$$

Thus, in view of (31), we have proved the second inequality in (3) with $L = \frac{1}{\kappa}$.

4 Gradient methods with inexact oracle

Consider the problem (1), where f is endowed with an inexact (δ, L)-oracle. In this section, we will use the Euclidean norm $\|.\|_2$. As usual when dealing with constrained problems, we assume that the gradient mapping $T_L(x,g)$ is computable for any $x \in Q$ and $g \in E^*$, see (8).

4.1 Primal gradient method

The classical (primal) gradient method can be adapted in a straightforward manner to accept first-order information from an inexact oracle: it is enough to replace the true gradient by its approximate counterpart $g_{\delta,L}$. Moreover,

we allow the parameters (δ_k, L_k) of the inexact oracle to be different for each iteration k. We obtain

Initialization: Choose $x_0 \in Q$.

Iteration $(k \geq 0)$: **1.** Choose δ_k and L_k.
2. Compute $(f_{\delta_k,L_k}(x_k), g_{\delta_k,L_k}(x_k))$.
3. Compute $x_{k+1} = T_{L_k}(x_k, g_{\delta_k,L_k}(x_k))$. (PGM)

Theorem 2 *For $k \geq 1$, we have*

$$\sum_{i=0}^{k-1} \tfrac{1}{L_i}[f(x_{i+1}) - f(x^*)] \leq \tfrac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k-1} \tfrac{\delta_i}{L_i}. \tag{32}$$

Proof Denote $r_k = \|x_k - x^*\|_2^2$, $f_k = f_{\delta_k,L_k}(x_k)$, and $g_k = g_{\delta_k,L_k}(x_k)$. Then

$$\begin{aligned} r_{k+1}^2 &= r_k^2 + 2\langle B(x_{k+1} - x_k), x_{k+1} - x^*\rangle - \|x_{k+1} - x_k\|_2^2 \\ &\overset{(11)}{\leq} r_k^2 + \tfrac{2}{L_k}\langle g_k, x^* - x_{k+1}\rangle - \|x_{k+1} - x_k\|_2^2 \\ &= r_k^2 + \tfrac{2}{L_k}\langle g_k, x^* - x_k\rangle - \tfrac{2}{L_k}[\langle g_k, x_{k+1} - x_k\rangle + \tfrac{L_k}{2}\|x_{k+1} - x_k\|_2^2] \\ &\overset{(3)}{\leq} r_k^2 + \tfrac{2}{L_k}[f(x^*) - f_k] - \tfrac{2}{L_k}[f(x_{k+1}) - f_k - \delta_k]. \end{aligned}$$

Summing up these inequalities for $i = 0, \ldots, k-1$, we obtain (32). □

When exact first-order information is used ($\delta_i = 0$, $L_i = L$), it is well-known that sequence $\{f(x_i)\}$ must be decreasing. This is no longer true when an inexact oracle is used. Therefore, let us define

$$\hat{x}_k = \tfrac{\sum_{i=0}^{k-1} L_i^{-1} x_{i+1}}{\sum_{i=0}^{k-1} L_i^{-1}} \in Q.$$

Since f is convex, we have

$$f(\hat{x}_k) - f(x^*) \leq \tfrac{\frac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k-1} L_i^{-1}\delta_i}{\sum_{i=0}^{k-1} L_i^{-1}}. \tag{33}$$

In the case when the oracle accuracy is constant ($\delta_i = \delta$, $L_i = L$), we have

$$f(\hat{x}_k) - f(x^*) \leq \tfrac{LR^2}{2k} + \delta, \quad R \overset{\text{def}}{=} \|x_0 - x^*\|_2. \tag{34}$$

Thus, there is no error accumulation, and the upper bound for the objective function accuracy decreases with k and asymptotically tends to δ. Hence, if an accuracy ϵ on the objective function is required (with $\epsilon > \delta$), $k = \tfrac{LR^2}{2(\epsilon-\delta)}$ iterations are sufficient. In particular, we see that PGM allows the oracle accuracy to be of the same order as the desired accuracy for the objective function.

4.2 Dual gradient method

This method [18] generates two sequences $\{x_k\}_{k\geq 0}$ and $\{y_k\}_{k\geq 0}$.

Initialization: Choose $x_0 \in Q$.

Iteration $(k \geq 0)$**:** **1.** Choose δ_k and L_k.

2. Compute $(f_{\delta_k,L_k}(x_k), g_{\delta_k,L_k}(x_k))$. (35)

3. Compute $x_{k+1} = \arg\min_{x\in Q}\left[\sum_{i=0}^{k} \frac{1}{L_i}\langle g_{\delta_i,L_i}(x_i), x - x_i\rangle + \frac{1}{2}\|x - x_0\|_2^2\right]$.

Define $y_k = T_{L_k}(x_k, g_{\delta_k,L_k}(x_k))$, $k \geq 0$.

Theorem 3 *For any $k \geq 0$ we have*

$$\sum_{i=0}^{k} \frac{1}{L_i}[f(y_i) - f(x^*)] \leq \frac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k} \frac{\delta_i}{L_i}. \tag{36}$$

Proof For $k \geq 0$, denote $f_k = f_{\delta_k,L_k}(x_k)$, $g_k = g_{\delta_k,L_k}(x_k)$, and

$$\psi_k(x) = \sum_{i=0}^{k} \frac{1}{L_i}[f_i + \langle g_i, x - x_i\rangle] + \frac{1}{2}\|x - x_0\|_2^2, \quad \psi_k^* = \min_{x\in Q} \psi_k(x).$$

In view of the first inequality in (3), we have for all $x \in Q$

$$\psi_k^* \leq \psi_k(x) \leq \sum_{i=0}^{k} \frac{1}{L_i} f(x) + \frac{1}{2}\|x - x_0\|_2^2. \tag{37}$$

Let us prove that $\psi_k^* \geq \sum_{i=0}^{k} \frac{1}{L_i}[f(y_i) - \delta_i]$. Indeed, this inequality is valid for $k = 0$:

$$f(y_0) \overset{(3)}{\leq} f_0 + \langle g_0, y_0 - x_0\rangle + \frac{L_0}{2}\|y_0 - x_0\|_2^2 + \delta_0 = L_0\psi_0^* + \delta_0.$$

Assume it is valid for some $k \geq 1$. Since $\Psi_k(x)$ is strongly convex, we have:

$$\psi_k(x) \geq \psi_k^* + \frac{1}{2}\|x - x_{k+1}\|_2^2, \quad x \in Q$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &= \min_{x\in Q}\left\{\psi_k(x) + \frac{1}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle]\right\} \\
&\geq \psi_k^* + \frac{1}{L_{k+1}} \min_{x\in Q}\left\{f_{k+1} + \langle g_{k+1}, x - x_{k+1}\rangle + \frac{L_{k+1}}{2}\|x - x_{k+1}\|_2^2\right\} \\
&\overset{(3)}{\geq} \psi_k^* + \frac{1}{L_{k+1}}(f(y_{k+1}) - \delta_{k+1}).
\end{aligned}$$

Hence, using our inductive assumption, we can prove that $\psi_k^* \geq \sum_{i=0}^{k} \frac{1}{L_i}[f(y_i) - \delta_i]$ for all $k \geq 0$. To conclude we combine this fact with inequality (37) for $x = x^*$. □

As for the Primal Gradient Method, we define

$$\hat{y}_k = \frac{\sum_{i=0}^{k} L_i^{-1} y_i}{\sum_{i=0}^{k} L_i^{-1}} \in Q,$$

and obtain the same upper bound

$$f(\hat{y}_k) - f(x^*) \leq \frac{\frac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k} L_i^{-1}\delta_i}{\sum_{i=0}^{k} L_i^{-1}}, \quad k \geq 0. \tag{38}$$

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

5 Fast gradient method with inexact oracle

5.1 Convergence analysis

In this section, we adapt one of the last versions of Fast Gradient Method (FGM) developed in [15]. Let $d(x)$ be a prox-function, differentiable and strongly convex 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

$$d(x_0) = 0, \quad d(x) \geq \tfrac{1}{2}\|x - x_0\|_E^2, \quad \forall x \in Q. \tag{39}$$

(here $\|\cdot\|_E$ denotes any norm on E). Let $\{\alpha_k\}_{k=0}^{\infty}$ be a sequence of reals such that

$$\alpha_0 \in (0, 1], \quad \frac{\alpha_k^2}{L_k} \leq A_k \stackrel{\text{def}}{=} \sum_{i=0}^{k} \frac{\alpha_i}{L_i}, \quad k \geq 0. \tag{40}$$

Define $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}L_{k+1}}$, $k \geq 0$ and consider the following method.

Initialization: Choose δ_0, L_0, and $x_0 = \arg\min_{x \in Q} d(x)$.

Iteration ($k \geq 0$): **1.** Compute $(f_{\delta_k, L_k}(x_k), g_{\delta_k, L_k}(x_k))$.

2. Compute $y_k = T_{L_k}(x_k, g_{\delta_k, L_k}(x_k))$. (FGM)

3. Compute $z_k = \arg\min_{x \in Q}\{d(x) + \sum_{i=0}^{k} \frac{\alpha_i}{L_i}\langle g_{\delta_i, L_i}(x_i), x - x_i\rangle\}$.

4. Choose δ_{k+1} and L_{k+1}. Define $x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k$.

Denote $\psi_k^* = \min_{x \in Q}\{d(x) + \sum_{i=0}^{k} \frac{\alpha_i}{L_i}[f_{\delta_i, L_i}(x_i) + \langle g_{\delta_i, L_i}(x_i), x - x_i\rangle]\}$.

Theorem 4 *For all $k \geq 0$, we have $A_k f(y_k) \leq \psi_k^* + E_k$ with $E_k = \sum_{i=0}^{k} A_i \delta_i$.*

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

$$\begin{aligned}\psi_0^* &= \min_{x \in Q} \Big\{ d(x) + \tfrac{\alpha_0}{L_0}[f_0 + \langle g_0, x - x_0 \rangle] \Big\} \\ &\overset{(39)}{\geq} \tfrac{\alpha_0}{L_0} \min_{x \in Q} \big\{ f_0 + \langle g_0, x - x_0 \rangle + \tfrac{L_0}{2} \|x - x_0\|_E^2 \big\} \overset{(3)}{\geq} \tfrac{\alpha_0}{L_0}[f(y_0) - \delta_0].\end{aligned}$$

Assume now that the statement of the theorem is true for some $k \geq 0$. Optimality conditions for the optimization problem solved at Step 3 imply

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

Hence, in view of strong convexity of d,

$$\begin{aligned} d(x) &\geq d(z_k) + \langle \nabla d(z_k), x - z_k \rangle + \tfrac{1}{2} \|x - z_k\|_E^2 \\ &\geq d(z_k) + \textstyle\sum_{i=0}^{k} \frac{\alpha_i}{L_i} \langle g_i, z_k - x \rangle + \frac{1}{2} \|x - z_k\|_E^2 . \end{aligned}$$

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

$$\begin{aligned} d(x) + \sum_{i=0}^{k+1} \tfrac{\alpha_i}{L_i}[f_i + \langle g_i, x - x_i \rangle] &\geq d(z_k) + \sum_{i=0}^{k} \tfrac{\alpha_i}{L_i}[f_i + \langle g_i, z_k - x_i \rangle] \\ &\quad + \tfrac{1}{2} \|x - z_k\|_E^2 + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle]. \end{aligned}$$

We have obtained

$$\psi_{k+1}^* \geq \psi_k^* + \min_{x \in Q} \{ \tfrac{1}{2} \|x - z_k\|_E^2 + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \}.$$

On the other hand, we have

$$\begin{aligned} &\psi_k^* + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ &\geq A_k f(y_k) - E_k + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ &\overset{(3)}{\geq} A_k[f_{k+1} + \langle g_{k+1}, y_k - x_{k+1} \rangle] - E_k + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}, x - x_{k+1} \rangle] \\ &= A_{k+1} f_{k+1} + \langle g_{k+1}, A_k(y_k - x_{k+1}) + \tfrac{\alpha_{k+1}}{L_{k+1}}(x - x_{k+1}) \rangle - E_k. \end{aligned}$$

Taking into account that

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

we obtain

$$\psi_k^* + \tfrac{\alpha_{k+1}}{L_{k+1}}[f_{k+1} + \langle g_{k+1}), x - x_{k+1})] \geq A_{k+1} f_{k+1} + \tfrac{\alpha_{k+1}}{L_{k+1}} \langle g_{k+1}, x - z_k \rangle - E_k.$$

Therefore,

$$\begin{aligned}
\psi_{k+1}^* &\geq A_{k+1} f_{k+1} - E_k + \min_{x \in Q} \{ \tfrac{1}{2} \|x - z_k\|_E^2 + \tfrac{\alpha_{k+1}}{L_{k+1}} \langle g_{k+1}, x - z_k \rangle \} \\
&= A_{k+1} \left[f_{k+1} + \min_{x \in Q} \{ \tfrac{1}{2A_{k+1}} \|x - z_k\|_E^2 + \tau_k \langle g_{k+1}, x - z_k \rangle \} \right] - E_k \\
&\overset{(40)}{\geq} A_{k+1} \left[f_{k+1} + \min\nolimits_{x \in Q} \{ \tfrac{\tau_k^2 L_{k+1}}{2} \|x - z_k\|_E^2 + \tau_k \langle g_{k+1}, x - z_k \rangle \} \right] - E_k.
\end{aligned}$$

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

$$\begin{aligned}
&\min_{x \in Q} \left\{ \tfrac{\tau_k^2 L_{k+1}}{2} \|x - z_k\|_E^2 + \tau_k \langle g_{k+1}, x - z_k \rangle \right\} \\
&= \min_{y} \left\{ \tfrac{L_{k+1}}{2} \|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1} \rangle : \ y \in \tau_k Q + (1 - \tau_k) y_k \right\} \quad (41) \\
&\geq \min_{y \in Q} \left\{ \tfrac{L_{k+1}}{2} \|y - x_{k+1}\|_E^2 + \langle g_{k+1}, y - x_{k+1} \rangle \right\}.
\end{aligned}$$

Therefore,we have:

$$\begin{aligned}
\Psi_{k+1}^* &\geq A_{k+1} \left[f_{k+1} + \min_{x \in Q} \{ \tfrac{\tau_k^2 L_{k+1}}{2} \|x - z_k\|_E^2 + \tau_k \langle g_{k+1}, x - z_k \rangle \} \right] - E_k \\
&\overset{(3),(41)}{\geq} A_{k+1} f(y_{k+1}) - E_k - A_{k+1} \delta_{k+1},
\end{aligned}$$

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

Theorem 5 *For all $k \geq 0$, we have $f(y_k) - f^* \leq \frac{1}{A_k} \left(d(x^*) + \sum_{i=0}^k A_i \delta_i \right)$.*

Proof Denote $f_i = f_{\delta_i, L_i}(x_i)$, and $g_i = g_{\delta_i, L_i}(x_i)$. Then

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

The proof now simply follows from the recurrence established in Theorem 4. □

A simple choice for the sequence $\{\alpha_i\}$ consists in letting $\alpha_i = \frac{i+1}{2}$. In that case, the sequence of Lipschitz constants must satisfy the inequality $\frac{(k+1)^2}{4L_k} \overset{(40)}{\leq} \sum\limits_{i=0}^{k} \frac{i+1}{2L_i}$, i.e.

$$L_k \geq \frac{(k+1)^2}{2} \Big/ \left[\sum_{i=0}^{k} \frac{i+1}{L_i}\right].$$

(It is true, for example, for any increasing sequence $\{L_k\}_{k\geq 0}$.) In this case, we obtain

$$f(y_k) - f^* \leq \frac{1}{\sum_{i=0}^{k} \frac{i+1}{2L_i}} \left(d(x^*) + \sum_{i=0}^{k} \sum_{j=0}^{i} \frac{j+1}{2L_j}\delta_i\right).$$

If the parameters of the inexact oracle are constant ($\delta_i = \delta$, $L_i = L$), we have $A_k = \frac{(k+1)(k+2)}{4L}$, $\tau_k = \frac{2}{k+3}$, and therefore

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)^2} + \frac{1}{(k+1)(k+2)} \sum_{i=0}^{k} (i+1)(i+2)\delta.$$

Since $\sum_{i=0}^{k}(i+1)(i+2) = \frac{1}{6}(k+1)(k+2)(2k+6)$, we obtain

$$f(y_k) - f^* \leq \frac{4Ld(x^*)}{(k+1)(k+2)} + \frac{1}{6}(2k+6)\delta \;=\; \frac{4LR^2}{(k+1)^2} + \frac{1}{3}(k+3)\delta \qquad (42)$$

where $R \overset{\text{def}}{=} \sqrt{d(x^*)}$.

5.2 Error accumulation

Contrarily to the classical gradient methods, the use of inexact oracle in FGM results in error accumulation. Indeed, while the first term in (42) decreases as $O(\frac{1}{k^2})$, the second term is increasing in k, and this FGM used with an inexact oracle is asymptotically divergent. Section 8.3 will prove that error accumulation and divergence are unavoidable for all fast first-order methods.

We now study the non-asymptotic behavior of FGM, and consider two cases.

a. Oracle accuracy δ is fixed. In this case, we can find the number of iterations k^* that achieves the minimal guaranteed residual for the objective function. We let

$$E(k) = \frac{4Ld(x^*)}{(k+1)^2} + \frac{1}{3}(k+1)\delta + \frac{2}{3}\delta.$$

This function is convex in k and its minimum is reached at iteration k^*

$$k^* = 2\sqrt[3]{\frac{3Ld(x^*)}{\delta}} - 1$$

for which the guaranteed accuracy for the objective function is

$$E(k^*) = \Theta(\delta^{2/3} L^{1/3} R^{2/3}).$$

b. Oracle accuracy δ can be chosen. Let us assume that parameter L of the inexact oracle is independent on δ. If we need to reach accuracy ϵ for the residual $f(y_k) - f^*$, it is enough to perform k iterations, with k satisfying two inequalities:

$$\tfrac{4Ld(x^*)}{(k+1)^2} \leq \tfrac{\epsilon}{2}, \quad \tfrac{1}{3}(k+3)\delta \leq \tfrac{\epsilon}{2}.$$

The first inequality gives us $k \geq \sqrt{\frac{8Ld(x^*)}{\epsilon}} - 1$, and the second one gives $k \leq \frac{3\epsilon}{2\delta} - 3$. Therefore attaining both $\frac{\epsilon}{2}$ accuracies is possible if and only if

$$\delta \leq \tfrac{3\epsilon^{3/2}}{2\sqrt{8Ld(x^*)}+4\sqrt{\epsilon}}. \tag{43}$$

In conclusion, if we choose the oracle accuracy satisfying relation (43), then after

$$k(\epsilon) = \sqrt{\tfrac{8Ld(x^*)}{\epsilon}} - 1$$

iterations, we obtain a point $y_{k(\epsilon)} \in Q$ satisfying $f(y_{k(\epsilon)}) - f^* \leq \epsilon$.

We observe that, compared to CGM, FGM requires a higher-order accuracy for the oracle ($\mathcal{O}(\epsilon^{3/2})$ versus $\mathcal{O}(\epsilon)$ for CGM).

6 Comparison between classical and fast gradient methods

When an exact oracle is used, FGM is an optimal method for the class $F_L^{1,1}(Q)$. It reaches an objective function accuracy ϵ after $O(\sqrt{\frac{L}{\epsilon}}R)$ iterations while CGM requires $O\left(\frac{LR^2}{\epsilon}\right)$ iterations for the same result.

Performing such a comparison becomes more complicated when an inexact first-order oracle is used. Contrary to CGM, FGM suffers from error accumulation. In order to compare their efficiency, we consider two cases.

6.1 Oracle accuracy δ can be freely chosen

In this case we assume that L is independent from the oracle accuracy δ (see examples in Section 3). If we need to reach ϵ accuracy for the objective function, CGM will work using an inexact oracle with $\delta = \Theta(\epsilon)$. However, it will then need $O\left(\frac{LR^2}{\epsilon}\right)$ iterations.

For FGM with inexact oracle, error accumulation forces the use of a more accurate oracle, i.e. with $\delta = \Theta\left(\frac{\epsilon^{3/2}}{\sqrt{L}R}\right)$. However only $O\left(\sqrt{\frac{L}{\epsilon}}R\right)$ iterations are needed. Thus, the choice between two methods depends on the complexity of the inexact oracle. Denote by $C(\delta)$ the cost associated with computing an answer $(f_{\delta,L}(x), g_{\delta,L}(x))$ for a (δ, L) inexact oracle. We see that CGM is preferable to FGM if the following holds (up to constant factors in the arguments of $C(\cdot)$)

$$\tfrac{1}{\epsilon}LR^2C(\epsilon) < \tfrac{1}{\epsilon^{1/2}}L^{1/2}RC\left(\tfrac{\epsilon^{3/2}}{L^{1/2}R}\right).$$

which leads us to consider the following situations.

- Oracle for which higher accuracy is very expensive: $C(\delta) = \Omega\left(\frac{1}{\delta}\right)$ (e.g. $C(\delta) = \frac{1}{\delta^2}$). In this case, it is preferable to use CGM.
- Oracle for which higher accuracy is moderately expensive: $C(\delta) = \Theta\left(\frac{1}{\delta}\right)$. For such an oracle both methods are equivalent.
- Oracle for which higher accuracy is cheap: $C(\delta) = o\left(\frac{1}{\delta}\right)$ (for example, $C(\delta) = \frac{1}{\delta^{1/2}}$, or even $C(\delta) = \ln\frac{1}{\delta}$). FGM is here better than CGM.

6.2 Oracle accuracy δ is fixed.

In this case, the sequence of iterates generated by CGM satisfies inequality

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

whereas the sequence obtained by FGM satisfies inequality

$$f(y^k) - f^* \leq \frac{4LR^2}{(k+1)(k+2)} + \frac{k+3}{3}\delta.$$

Figures 3, 4 and 2 depict these two rates of convergence for three different values of the oracle accuracy parameter δ (with $L = R = 1$ in all cases).

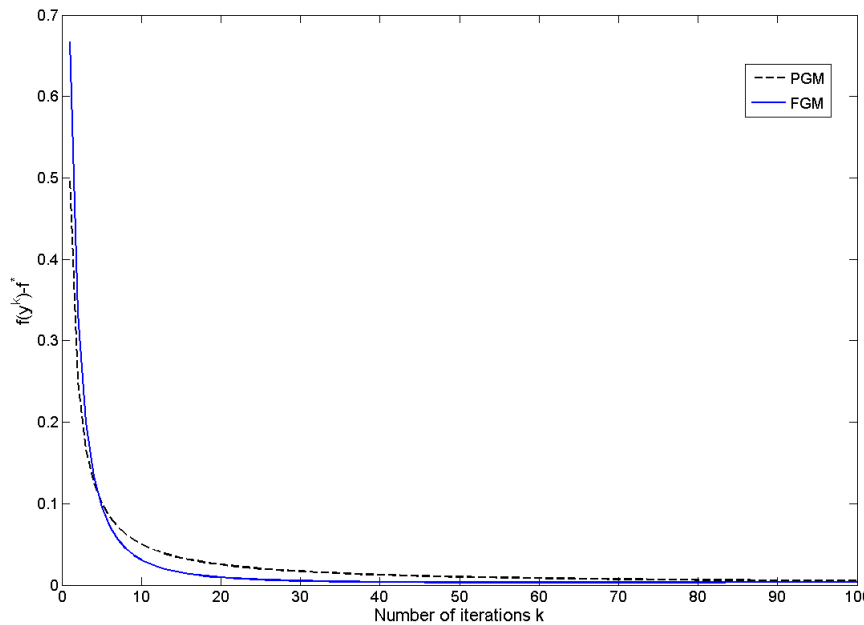

Fig. 2 Convergence rate of CGM and FGM with $\delta = 0.0001$, $L = 1$ and $R = 1$

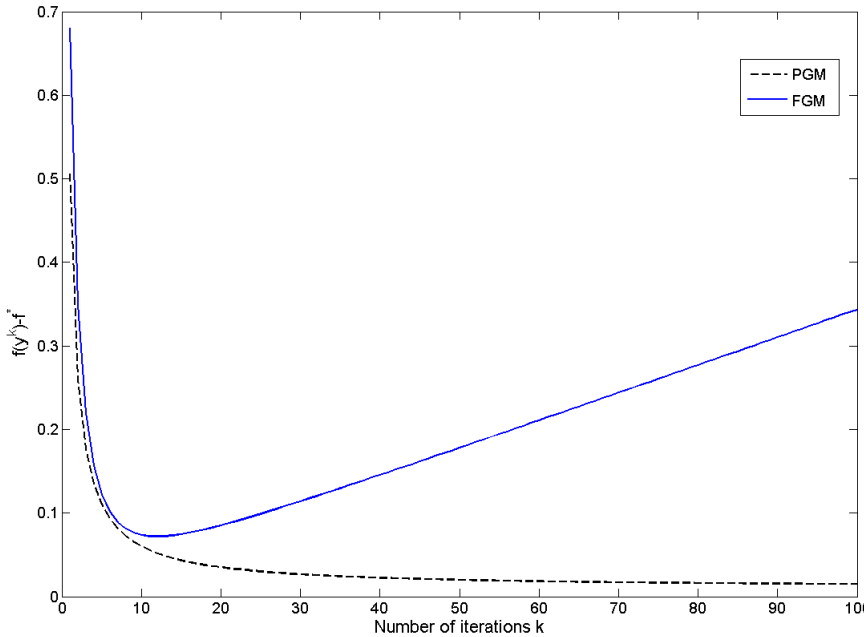

Fig. 3 Convergence rate of CGM and FGM with $\delta = 0.01$, $L = 1$ and $R = 1$

The higher the accuracy of the oracle, the larger the number of iterations for which FGM is better than CGM. For example, on Figure 2, we see that when the oracle accuracy is sufficiently high ($\delta = 0.0001$), FGM outperforms GM accuracy at least for the first hundred iterations (except for a few initial iterations, where smaller constant factors benefit CGM). In the exact case, i.e. when the oracle accuracy $\delta = 0$, FGM outperforms CGM for any number of iterations.

On the other hand, when oracle accuracy is low, accumulation of oracle errors in FGM becomes so prevalent that CGM is always better than FGM. Figure 3 (with $\delta = 0.01$) depicts this situation.

For intermediate values of accuracy (such as on Figure 4, where $\delta = 0.001$), the situation is more complicated. Better constant factors in the convergence rate initially lead to smaller errors for the first few iterations of CGM. After that, FGM reduces errors much better than CGM, because of its better convergence rate. For FGM, the error attains its minimum value after $N_1 = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ iterations, with corresponding accuracy $\delta^* = \Theta(\delta^{2/3}L^{1/3}R^{2/3})$. It is not interesting to perform further FGM iterations since the gap can then only increase due to error accumulation.

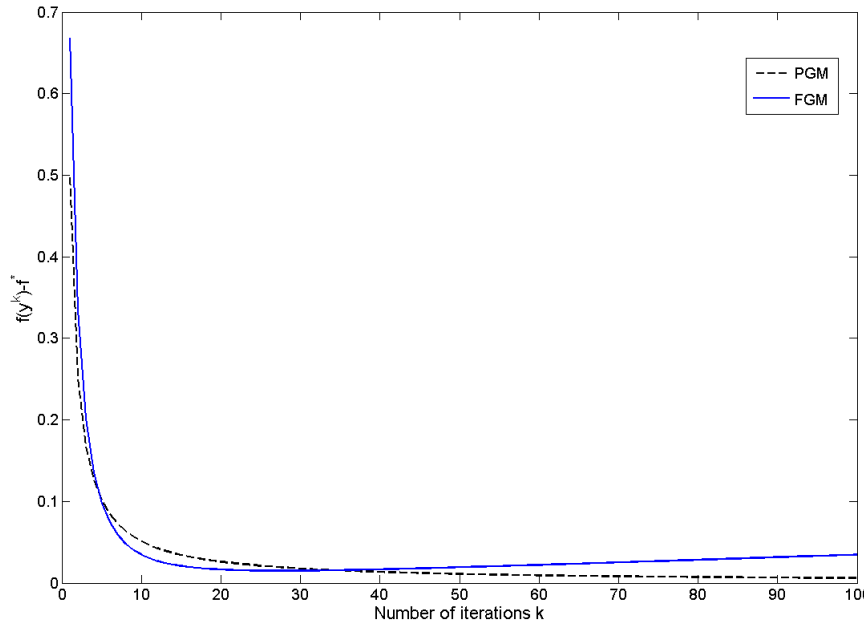

Fig. 4 Convergence rate of CGM and FGM with $\delta = 0.001$, $L = 1$ and $R = 1$

Note that there exists an iteration threshold N_2 ($> N_1$) after which CGM provides better accuracy than FGM. However, this does not mean that CGM is superior to FGM as soon as we reach that number of iterations, because FGM already achieved a lower accuracy δ^* after N_1 iterations. If we wait further until we reach $N_3 = \Theta\left(\frac{LR^2}{\delta^{2/3}}\right)$ ($> N_2$) iterations, the accuracy of CGM finally becomes better than δ^*, the best reachable accuracy with FGM. Final accuracies ϵ between δ^* and δ can then only be reached by CGM (they are inaccessible by FGM), and require $\Theta\left(\frac{LR^2}{\epsilon-\delta}\right)$ iterations.

In conclusion, FGM is the method of choice when we need accuracy not better than $\delta^* = \Theta(\delta^{2/3}L^{1/3}R^{2/3})$. Indeed, accuracy δ^* is reached by the FGM after N_1 iterations whereas the CGM needs N_3 iterations in order to obtain the same error. In order to obtain accuracy better than δ^*, CGM must be used since the FGM cannot decrease the error below δ^*.

7 Comparison with other types of inexact oracle

Fast-gradient methods using inexact first-order oracle have been recently studied in [3] and [1]. These works assume that set Q is bounded and that the oracle

provides at each point $y \in Q$ an approximate gradient $g(y)$ satisfying condition

$$|\langle g(y) - \nabla f(y), x - z\rangle| \leq \xi \quad \forall x, y, z \in Q. \tag{44}$$

Let us compare this definition with (3), taking into account both their applicability and the results obtained. First of all, the existence of an inexact oracle satisfying (44) require more assumptions than our definition:

- Set Q must be bounded (this is not needed for (3)).
- Objective function f must be differentiable. The existence of the gradient at all points is necessary since it must be compared with the approximate gradient. Our approach is also able to consider non- or weakly smooth convex functions.

Furthermore, even in the smooth case $f \in F_L^{1,1}(Q)$ with bounded Q, we argue that condition (44) is strictly stronger than (3). Assume $f \in F_L^{1,1}(Q)$.

1. Any approximate gradient $g(y)$ satisfying (44) also satisfies our definition. Indeed, in view of (2) and (44), we have for all $x, y \in Q$

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

   and therefore taking $f_{\delta,L}(y) = f(y) - \xi$, and $g_{\delta,L}(y) = g(y)$ satisfies (3) with $\delta = 2\xi$ and the same value for L.
2. On the other hand, our condition (3) does not imply (44) with any $\xi = \Theta(\delta)$. Indeed, consider the function $f(x) = \max_{u\in U} \Psi(x, u)$, where

   $$\Psi(x, u) = -\tfrac{1}{2}\|u\|_2^2 + \langle x, u\rangle, \quad Q = \{y \in \mathbb{R}^n : \|y\|_2 \leq 1\}, \quad U = \mathbb{R}^n. \tag{45}$$

   Let us assume the answer of oracle for $x = 0$ is obtained for some point u_0 satisfying $\|u_0\|_2 = \delta^{1/2}$. Since $u_0^* = \arg\max_{u\in U} \Psi(0, u) = 0$, and $f(0) - \Psi(0, u_{x_0}) = \frac{1}{2}\|u_{x_0}\|_2^2 = \frac{\delta}{2}$, the pair $(f_{\delta,L}(0), g_{\delta,L}(0)) = (-\frac{\delta}{2}, u_0)$ is an acceptable answer for a (δ, L) inexact oracle with $L = 2$ (see Section 3.2). However we can check that

   $$\max_{y,z\in Q} |\langle \nabla f(0) - g_{\delta,L}(0), y - z\rangle| = 2\max_{y\in Q} |\langle u_0, y\rangle| = 2\delta^{1/2}.$$

We now compare efficiency estimates of FGM based on these oracles. FGM using oracle (44) converges as follows:

$$f(y_k) - f^* \leq \tfrac{CLR^2}{k^2} + 3\xi,$$

where C is an absolute constant. This bound does not feature error accumulation, meaning the accuracy ξ of the oracle can be chosen to be of the same order as the desired accuracy ϵ of the solution. This result seems at first sight to be better than what we obtained with our (δ, L)-oracle.

However, we noted that for the same level of accuracy, condition (44) is much stronger than (3). Let us look at important example. Consider the class of functions with explicit max-structure: $f(x) = \max_{u\in U} \Psi(x, u)$, where set U is closed and convex, and $\Psi(x, u) = G(u) + \langle x, Au\rangle$, where $G(u)$ is a differentiable,

strongly concave function with concavity parameter κ. Assume that we want to solve the primal problem $\min_{x\in Q} f(x)$ with accuracy ϵ. With our definition of inexact oracle, the oracle accuracy δ corresponds directly to the (objective function) accuracy required when solving the dual problem (see Section 3.2).

In the case of an approximate gradient satisfying definition (44), we can also use an approximate dual solution $u_x \approx u_x^*$

$$\nabla f(x) = Au_x^*, \quad g(x) = Au_x.$$

However, we need to satisfy the following relation:

$$|\langle A(u_x^* - u_x), y - z\rangle| \leq \epsilon, \quad \forall x, y, z \in Q. \tag{46}$$

(We can take $\xi = \epsilon$ since the condition (44) avoids accumulation of errors). For that, we need to have u_x close to u_x^* according to

$$\|u_x - u_x^*\|_F \leq \frac{\epsilon}{\operatorname{diam}(Q)\cdot\|A\|_{F\to E^*}}.$$

Since Ψ is strongly concave, i.e. $\Psi(x, u_x^*) - \Psi(x, u_x) \geq \frac{\kappa}{2}\|u_x - u_x^*\|_F^2$, a sufficient condition for (44) is then as follows

$$\Psi(x, u_x^*) - \Psi(x, u_x) \leq \frac{\kappa}{2}\left(\frac{\epsilon}{\operatorname{diam}(Q)\cdot\|A\|_{F\to E^*}}\right)^2 = \mathcal{O}(\epsilon^2).$$

Compare this to our approach, for which it was enough to solve the dual problem up to accuracy $\epsilon^{3/2}$ (see (43)) in order to avoid accumulation of errors.

Remark 2 In some cases, inequality $\Psi(x, u_x^*) - \Psi(x, u_x) \leq \epsilon^2/8$ is also a necessary condition for (46). Indeed, consider again the saddle point problem defined by (45). We have $f(0) - \Psi(0, u_0) = \frac{1}{2}\|u_0\|_2^2$. In order to satisfy condition (46) we need to ensure

$$\epsilon \geq 2\max_{y\in Q}|\langle u_0, y\rangle| \;=\; 2\|u_0\|_2 = 2\sqrt{2(f(0) - \Psi(0, u_0))}.$$

Remark 3 The definition of inexact oracle used in [1] is slightly different from (44). The author assume that $g(y)$ satisfies the following conditions:

$$f(x) \geq f(y) + \langle g(y), x - y\rangle - \bar{\xi} \quad \forall x \in \operatorname{dom} f$$
$$f(x) \geq f(y) + \langle g(y), x - y\rangle - \bar{\xi}\|x - y\| \quad \forall x \in \operatorname{dom} f$$

and that the set Q is bounded. It is possible to prove that this definition implies (44) with $\xi = D_Q\bar{\xi}$ (where D_Q denotes the diameter of Q), possibly replacing $\nabla f(y)$ with a subgradient when function f is non-smooth.

8 Applications to non-smooth optimization

8.1 Solving weakly smooth problems

Let f be a convex function satisfying the Hölder condition (14). This class includes non-smooth convex functions with bounded variation of subgradients ($\nu = 0$), and smooth convex functions with Hölder continuous gradient ($\nu \in (0,1]$). We have shown in Section 2, that for all $\delta > 0$ these functions can be equipped with (δ, L)-oracle with

$$L = A(\delta,\nu) \;=\; L_\nu \left[\tfrac{L_\nu}{2\delta}\cdot\tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}.$$

This observation allows us to apply first-order methods of $F_L^{1,1}(Q)$ to functions with weaker level of smoothness, replacing the gradients by subgradients and using a Lipschitz constant L that grows as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$ in terms of the δ parameter of the oracle.

This parameter δ, which does not correspond to the actual accuracy of the oracle, will have to be properly tuned in numerical methods, with a tradeoff between the high "accuracy" of the oracle, and a small Lipschitz constant L.

For the sake of simplicity, we assume in the rest of this section that a fixed number of iterations N is performed.

Let us apply method PGM to a weakly smooth function f with an inexact (δ, L)-oracle. In view of (34), after N iterations we have

$$f(\hat{x}_N) - f(x^*) \le L_\nu \left[\tfrac{L_\nu}{2\delta}\cdot\tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{2N} + \delta \;\stackrel{\text{def}}{=}\; C_N \left(\tfrac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} + \delta.$$

Denote $\tau = \frac{1-\nu}{1+\nu}$. Then the optimal accuracy δ_N can be found from the equation

$$C_N \tfrac{\tau}{\delta_N^{1+\tau}} = 1.$$

Thus, we come to the following bound:

$$f(\hat{x}_N) - f(x^*) \le \delta_N \left(\tfrac{C_N}{\delta_N^{1+\tau}} + 1\right) \;=\; \tfrac{2\delta_N}{1-\nu}. \tag{47}$$

Note that

$$\delta_N = (\tau C_N)^{\frac{1}{1+\tau}} \;=\; \left(\tfrac{1-\nu}{1+\nu}\cdot L_\nu \left[\tfrac{L_\nu}{2}\cdot\tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{2N}\right)^{\frac{1+\nu}{2}} \;=\; \tfrac{1-\nu}{1+\nu}\cdot \tfrac{L_\nu R^{1+\nu}}{2^{\frac{1-\nu}{2}}\cdot N^{\frac{1+\nu}{2}}}.$$

Thus, we come to the following upper bound:

$$f(\hat{x}_N) - f(x^*) \le \tfrac{L_\nu R^{1+\nu}}{1+\nu}\cdot\left(\tfrac{2}{N}\right)^{\frac{1+\nu}{2}}. \tag{48}$$

For functions with bounded variation of subgradients ($\nu = 0$), we get:

$$f(\hat{x}_N) - f(x^*) \le L_0 R\cdot\left(\tfrac{2}{N}\right)^{\frac{1}{2}},$$

which is the optimal rate of convergence (see [11,14]). However for functions with Hölder continuous gradient ($0 < \nu$), the obtained rate is not optimal (it should be $O(N^{-\frac{1+3\nu}{2}})$, see [10,8]).

Let us now apply FGM to a weakly smooth function using an (δ, L)-oracle. In view of (42), after N iterations we have:

$$f(y_N) - f(x^*) \leq 4L_\nu \left[\frac{L_\nu}{2\delta} \cdot \frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \frac{R^2}{(N+1)^2} + \delta \cdot (N+1)$$

$$\stackrel{\text{def}}{=} \hat{C}_N \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} + \delta \cdot (N+1).$$

The equation for optimal δ_N now becomes $\hat{C}_N \frac{\tau}{\delta_N^{1+\tau}} = N+1$. Therefore, we get

$$f(y_N) - f(x^*) \leq \delta_N \left(\frac{\hat{C}_N}{\delta_N^{1+\tau}} + N + 1\right) = \frac{2\delta_N}{1-\nu}(N+1).$$

Note that

$$\delta_N = (\hat{C}_N \tfrac{\tau}{N+1})^{\frac{1}{1+\tau}} = \left(\frac{1-\nu}{1+\nu} \cdot 4L_\nu \left[\frac{L_\nu}{2} \cdot \frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}} \frac{R^2}{(N+1)^3}\right)^{\frac{1+\nu}{2}}$$

$$= \frac{1-\nu}{1+\nu} \cdot \frac{L_\nu R^{1+\nu}}{(N+1)^{\frac{3}{2}(1+\nu)}} \cdot 2^{\frac{1+3\nu}{2}}.$$

Thus, we obtain the following upper bound

$$f(y_N) - f(x^*) \leq \frac{2L_\nu R^{1+\nu}}{1+\nu} \left(\frac{2}{N+1}\right)^{\frac{1+3\nu}{2}}. \tag{49}$$

For functions with bounded variation of subgradients ($\nu = 0$), we get

$$f(y_N) - f(x^*) \leq 2L_0 R \left(\frac{2}{N+1}\right)^{\frac{1}{2}}.$$

In all cases, we obtain the optimal rate of convergence. Therefore, FGM can be seen as a universal first-order method simultaneously optimal for smooth, weakly smooth and non-smooth convex functions.

The applicability of first-order method of smooth convex optimization to non-smooth convex problems, justified by the notion of (δ, L)-oracle, has several further interesting consequences. We describe two of them below.

- We can apply CGM and FGM to objective functions composed of a sum of smooth and non-smooth components.
- We can get lower bounds on the rate of accumulation of errors in the first-order methods based on (δ, L)-oracle. It appears that error accumulation is an intrinsic property of any fast gradient method. Slower first-order methods can avoid accumulation of errors, and CGM is the fastest method among those.

8.2 Solving composite optimization problems

Consider the composite convex objective function:

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

where f_1 is a smooth convex function with Lipschitz continuous gradient (constant $L(f_1)$), and f_2 is a non-smooth convex function with subgradients whose variation is bounded by constant $M(f_2)$. We assume that the standard exact first-order oracles are available for both f_1 and f_2.

Note that function f_1 is equipped with $(0, L(f_1))$-oracle, and by (17) function f_2 has $(\delta, \frac{1}{2\delta}M^2(f_2))$-oracle. Hence, we conclude that the pair

$$(f_1(y) + f_2(y), \nabla f_1(y) + g_2(y)), \quad g_2(y) \in \partial f_2(y), \tag{50}$$

is a (δ, L)-oracle for function f with $L = L(f_1) + \frac{1}{2\delta}M^2(f_2)$. Assume again that the number of iterations N for our methods is fixed.

Let us apply now CGM to function f using the inexact (δ, L)-oracle (50). Then, after N iterations we have:

$$f(\hat{x}_N) - f^* \overset{(34)}{\leq} \left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right)\tfrac{R^2}{2N} + \delta.$$

Minimizing this expression with respect to $\delta \geq 0$, we obtain $\delta^* = \frac{M(f_2)R}{2N^{1/2}}$. Therefore, the best upper bound for the residual is

$$f(\hat{x}_N) - f^* \leq \tfrac{L(f_1)R^2}{2N} + \tfrac{M(f_2)R}{N^{1/2}}.$$

This method has the optimal rate of convergence for non-smooth part of the problem, but not for the smooth one.

Let us check now the performance of FGM as applied to the composite problems. In view of (42), we have after N iterations of the scheme

$$f(y_N) - f^* \leq 4\left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right)\tfrac{R^2}{(N+1)^2} + \delta \cdot (N+1).$$

Minimizing this function in $\delta \geq 0$, we obtain: $\delta^* = \frac{2^{1/2}M(f_2)R}{(N+1)^{3/2}}$. The upper-bound therefore becomes

$$f(y_N) - f^* \leq \tfrac{4L(f_1)R^2}{(N+1)^2} + \tfrac{2^{3/2}M(f_2)R}{(N+1)^{1/2}}.$$

For such a composite objective function, this method is optimal both for the smooth and non-smooth parts of the problem.

Remark 4 Our analysis is in a certain sense similar to that of [5], where the author applies a version of FGM to a stochastic composite optimization problem.

In the deterministic case, the author applies a variant of FGM, replacing the gradients by subgradients in the non-smooth part of objective, and the Lipschitz constant by a quantity of order $\mathcal{O}(M(f_2)N^{3/2})$. This method appears

to be optimal both for the smooth and non-smooth parts of the composite function.

In our approach, $N = \Theta((\frac{1}{\delta}M(f_2))^{2/3})$, and we get $M(f_2)N^{3/2} = \Theta(\frac{1}{\delta}M^2(f_2))$, which is, up to a constant factor, the quantity that replaces the Lispchitz constant for our method.

8.3 First-order methods and error accumulation

Applicability of first-order methods of smooth optimization to non-smooth problems, based on the notion of inexact oracle, opens a possibility to derive lower bounds on error accumulation. This is the main subject of this section. We start from the following observation.

Theorem 6 *Consider a first-order method for $F_L^{1,1}(Q)$ with convergence rate $O(\frac{LR^2}{k^p})$ when exact first-order information is used. Assume that the bounds on the performance of this method, as applied to a problem equipped with an inexact (δ, L)-oracle, are given by inequality*

$$f(z_k) - f^* \leq \frac{C_1 L R^2}{k^p} + C_2 k^q \delta, \tag{51}$$

where C_1, C_2 are absolute constants, and k is the iteration counter. Then the inequality $q \geq p - 1$ must hold.

Proof Let f be a non-smooth convex function, whose subgradients have variation bounded by constant M. We have seen that for such a function, the standard oracle can be treated as $(\delta, \frac{M^2}{2\delta})$-oracle for any $\delta > 0$. Therefore, by our method we can ensure the following rate of convergence:

$$f(z_k) - f^* \leq \frac{C_1 M^2 R^2}{2\delta k^p} + C_2 k^q \delta.$$

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

$$f(z_k) - f^* \leq [2C_1 C_2]^{1/2} M R \cdot k^{-\frac{p-q}{2}}.$$

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

In the exact case, when minimizing a function in $F_L^{1,1}(Q)$, any first-order method with convergence rate $\Theta(\frac{LR^2}{k^2})$ is optimal (e.g. FGM), and any method with the convergence rate $\Theta(\frac{LR^2}{k})$ is suboptimal (e.g. CGM). In the case of inexact (δ, L)-oracle, the situation is more complicated.

Total performance of the method also depends from the way it accumulates successive errors coming from the oracle. In this situation, the superiority of FGM over CGM is not completely clear anymore. As we have seen in the previous sections, FGM suffers from accumulation of errors, but CGM does not.

From Theorem 6, we know that this accumulation is a direct consequence of the fast convergence of the scheme. Any method with complexity estimate

$\Theta(\sqrt{\frac{L}{\epsilon}}R)$ must suffer from this instability. On the other hand, it appears that in the inexact situation, both FGM and CGM are optimal, but in different senses.

- $q = 0 \Rightarrow p \leq 1$:
  It is impossible to have a first-order method without accumulation of errors, which has better complexity than CGM, that is $\Theta(\frac{LR^2}{\epsilon})$.
- $p = 2 \Rightarrow q \geq 1$:
  On the other hand, if we have a first-order method with complexity $\Theta(\sqrt{\frac{L}{\epsilon}}R)$, then it always has accumulating of errors, which grow at least as $\Theta(k\delta)$.

The next theorem relates the rate of convergence of the method with the required accuracy of the oracle.

Theorem 7 *Let parameter L of inexact oracle (3) be independent from δ. Under assumptions of Theorem 6, accuracy ϵ in the residual of objective function requires at least the following accuracy of the oracle:*

$$\delta \leq \frac{p\cdot\epsilon}{(p+q)C_2}\left[\frac{q\cdot\epsilon}{(p+q)C_1LR^2}\right]^{q/p}.$$

Proof In order to guarantee accuracy ϵ by the estimate (51), we have to choose k and δ such that:

$$\frac{C_1LR^2}{k^p} \leq \alpha\epsilon, \quad C_2k^q\delta \leq (1-\alpha)\epsilon$$

for some $\alpha \in [0,1]$. The first inequality gives us $k \geq \left[\frac{C_1LR^2}{\alpha\epsilon}\right]^{1/p}$, and using the second inequality, we obtain

$$C_2\left[\frac{C_1LR^2}{\alpha\epsilon}\right]^{q/p}\delta \leq (1-\alpha)\epsilon.$$

Thus, $\delta \leq \frac{(1-\alpha)\alpha^{q/p}\cdot\epsilon^{(p+q)/p}}{C_2[C_1LR^2]^{q/p}}$. It remains to maximize the right-hand side of this inequality in α. □

Corollary 2 *If a first-order method has efficiency estimate $\Theta\left(\frac{LR^2}{\epsilon}\right)$, then it can be applied to an (δ, L)-oracle, with accuracy at least $\Omega(\frac{\epsilon^{1+q}}{L^qR^{2q}})$ or higher. For the method optimal with respect to accumulation of errors ($q = p-1 = 0$), we can choose $\delta = \Omega(\epsilon)$.*

Corollary 3 *If a first-order method has efficiency estimate $\Theta\left(\sqrt{\frac{L}{\epsilon}}R\right)$, then it can be applied to an (δ, L)-oracle, with accuracy at least $\Omega(\frac{\epsilon^{1+q/2}}{L^{q/2}R^q})$ or higher. For the method optimal with respect to accumulation of errors ($q = p-1 = 1$), we can choose $\delta = \Omega(\frac{\epsilon^{3/2}}{L^{1/2}R})$.*

References

1. M. Baes. Estimate sequence methods: extensions and approximations. *IFOR Internal report, ETH Zurich, Switzerland*, (2009)
2. R. Correa and C. Lemarechal. Convergence of some algorithms for convex minimization. *Mathematical Programming, Serie A*,**62**, 261-275 (1993).
3. A. D'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal of Optimization*, **19**, 1171-1183 (2008).
4. O. Devolder, F. Glineur and Y. Nesterov. Double smoothing technique for infinite-dimensional optimization problems with applications to optimal control. *CORE Discussion Paper*, **34**, (2010)
5. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming Serie A, Online First* (2010)
6. M. Hintermuller. A proximal bundle method based on approximative subgradient. *Computational Optimization and Applications*, **20**, 245-266 (2001)
7. K. Kiwiel. A proximal bundel method with approximative subgradient linearization. *SIAM Journal of Optimization*, **16**, 1007-1023 (2006)
8. L. Kachiyan, A Nemirovskii and Y. Nesterov. Optimal methods of convex programming and polynomial methods of linear programming. *In H. Elster, editor, Modern Mathematical Methods of Optimization*, Akademie Verlag 75-115 (1993).
9. A. Nedic and D.Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical programming, Serie A*, **125**, 75-99 (2010).
10. A. Nemirovskii and Y.Nesterov. Optimal methods for smooth convex minimization. *Zh. Vichisl. Mat. Fiz. (In Russian)*, **25(3)**, 356-369 (1985).
11. A. Nemirovskii and D. Yudin.Problem complexity and method efficiency in optimization. *John Wiley* (1983)
12. 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).
13. 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).
14. Yu. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers (2004)
15. Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming, Serie A*, **103**, 127-152 (2005).
16. Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *Siam Journal of Optimization*, **16**, 235-249 (2005).
17. Yu. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming A*, **110**, 245-259 (2007).
18. Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper*, **76**, (2007)
19. B.T. Polyak. Introduction to Optimization. *Optimization Software Inc* (1987)
20. N.Z. Shor. Minimization Methods for Non-Differentiable Functions. *Springer Series in Computational Mathematics. Springer-Verlag* (1985).
21. P. Tseng. On accelerated Proximal Gradient Methods for Convex-Concave Optimization *Submitted to Siam. J. Optim.* (2008).