2011/2

First-order methods of smooth convex optimization with inexact oracle

Olivier Devolder, François Glineur
and Yu. Nesterov

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

CORE DISCUSSION PAPER
2011/2

First-order methods of smooth convex optimization with inexact oracle

Olivier DEVOLDER [1], François GLINEUR[2]
and Yu. NESTEROV [3]

January 2011

Abstract

In this paper, we analyze different first-order methods of smooth convex optimization employing inexact first-order information. We introduce the notion of an approximate first-order oracle. The list of examples of such an oracle includes smoothing technique, Moreau-Yosida regularization, Modified Lagrangians, and many others. For different methods, we derive complexity estimates and study the dependence of the desired ac- curacy in the objective function and the accuracy of the oracle. It appears that in inexact case, the superiority of the fast gradient methods over the classical ones is not anymore absolute. Contrary to the simple gradient schemes, fast gradient methods necessarily suffer from accumulation of errors. Thus, the choice of the method depends both on desired accuracy and accuracy of the oracle. We present applications of our results to smooth convex-concave saddle point problems, to the analysis of Modified Lagrangians, to the prox-method, and some others.

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

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: Olivier.devolder@uclouvain.be.

[2] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: françois.glineur@uclouvain.be. This author is also member of ECORE, the association between CORE and ECARES.

[3] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: yurii.nesterov@uclouvain.be. This author is also member of ECORE, the association between CORE and ECARES.

The first author is a F.R.S.-FNRS Research Fellow.

This paper 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 is assumed by the authors.

1 Introduction

In large-scale convex optimization, first-order methods remain the methods of choice due to their cheap iteration cost. When the objective function is assumed to be smooth (e.g. its gradient is Lipschitz-continuous with constant L), the simplest numerical schemes to be considered are the gradient method and its variants. However, it is well known that these methods exhibit non optimal worst-case complexity of $O\left(\frac{L}{\epsilon}\right)$ iterations, where ϵ is the desired accuracy for the objective function.

In the black-box framework, the first-order methods that achieve the lower complexity bound of $O\left(\sqrt{\frac{L}{\epsilon}}\right)$ iterations, have been developed for various classes of problems since 1983 [18, 19, 13, 14]. Theses schemes, also called Fast Gradient Methods (FGM), outperform theoretically, and often in practice, the classical gradient methods. A new interest to this field has appeared in the last years with development of smoothing technique for non-smooth convex problems (see [14, 15, 16, 4]). In this approach, FGMs are used for minimizing a smooth approximation of the initial nonsmooth objective function.

All these first-order methods need an exact first-order information. Namely, at each point, the oracle must provide an exact value of the function and its gradient. However, in the problem obtained by the smoothing technique, the gradient of the modified objective function is computed by solving another auxiliary optimization problem. In many situations in practice, we are able to solve this subproblem only approximately. Hence, the first-order information given to numerical methods is often inexact. This is only one among many other examples, which motivate our research in analyzing the behavior of first-order methods working with inexact oracle.

In this paper, in Section 2 we introduce a new definition of inexact first-order oracle and give some simple examples. In Section 3, we show how our concept works in the situations when the inexact oracle is computed by an auxiliary optimization problem. In particular, we consider convex-concave saddle point problems, modified Lagrangians, and Moreau-Yosida regularization.

In the Sections 4 and 5, we look at the classical and fast gradient methods for $F_L^{1,1}(Q)$, the class of convex functions which gradient is Lipschitz-continuous on convex set Q with constant L. We obtain their efficiency estimates under assumption that the available oracle provides us only with approximative first-order information. For each method, we also study the link between the desired accuracy in the objective function and the necessary accuracy of the oracle.

It appears that in inexact case, the superiority of FGM over the classical one is not anymore absolute. If the accuracy of the oracle is not high enough, any FGM, contrarily to the classical gradient method, suffers from accumulation of errors. Hence, the choice between these methods depends on the relative complexity of computations in inexact oracle. This comparison is done in Section 6.

In Section 7, we compare our approach with other popular 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.

In Section 8, we discuss the consequences of the applicability of our definition of inexact oracle to non-smooth and weakly-smooth convex problems. In our approach, it is possible to apply any first-order method of smooth convex optimization (i.e. developed for the class $F_L^{1,1}(Q)$) to functions with a weaker level of smoothness. For that, we just replace in the method the gradients by subgradients (this is for non-smooth case), and use the Lipschitz constants, which grow with the desired accuracy. In this way, we can obtain a "universal" first-order method, which has the *optimal* rate of convergence for objective functions with different level of smoothness. By this application, we prove the *lower bounds* on the rate of accumulation of errors in the first-order methods. It appears that accumulation of errors is a intrinsic property of any FGM. The slower gradient methods are able to keep the error on the level of accuracy of the oracle. All methods discussed in our paper have the lowest possible rate of accumulation of errors.

In the last Section 9, for the problems with strongly convex objective function, we obtain the complexity results and study the links between oracle accuracy and desired accuracy for the solution.

2 Definition of inexact first-order oracle

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. The space E is endowed with the norm $\|\cdot\|_E$ and E^*, the dual space of E, with $\|g\|_E^* = \sup_{y\in E}\{|\langle g, y\rangle| : \|y\|_E \leq 1\}$ where $\langle .,. \rangle$ denotes the dual pairing. Let (1) be solvable with optimal solution x^*.

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 for all $x \in Q$ we have*

$$\begin{aligned} f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle \;&\leq\; f(x) \\ \leq\; f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle &+ \tfrac{L}{2}\|x-y\|_E^2 + \delta. \end{aligned} \tag{2}$$

We denote by $\mathcal{O}_{\delta,L}[f](y) = (f_{\delta,L}(y), g_{\delta,L}(y))$ the response of the oracle at point y.

In some applications, the Lipschitz constant L is a *function* of the oracle accuracy δ, which can be chosen arbitrarily. In this case, we have a one-parametric family of $(\delta, L(\delta))$-oracles.

Recall that for functions in $F_L^{1,1}(Q)$, for any pair of point $x, y \in Q$ we have

$$f(y) + \langle\nabla f(y), x-y\rangle \leq f(x) \leq f(y) + \langle\nabla f(y), x-y\rangle + \tfrac{L}{2}\|x-y\|_E^2. \tag{3}$$

Thus, our definition is a generalization of the properties of the standard first-order oracle providing the exact gradient and the exact function value. However, as we will see soon, our approach is not restricted by the functions from $F_L^{1,1}(Q)$.

Let us mention the most important properties of (δ, L)-oracle.

- Taking in (2) $x = y$, we obtain:

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

  Thus, $f_{\delta,L}(y)$ is a lower δ-approximation of the function value.

- For all $x, y \in Q$ we have

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

  Therefore $g_{\delta,L}(y)$ is an δ-subgradient of f at $y \in Q$:

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

  Methods of non-smooth convex optimization based on δ-subgradients have a long history (see e.g. [21, 20, 2, 10] for subgradient methods, and [2, 7, 8] for proximal point and bundle methods). In our paper, we will show that the second inequality in (2) can be satisfied even by usual subgradient. This opens a possibility for using FGM in nonsmooth convex optimization.

- If $\langle g_{\delta,L}(y), x - y\rangle \geq 0$, for all $x \in Q$, then $f_{y,\delta} \leq f^*$ and therefore $f(y) \leq f^* + \delta$. Thus, (δ, L) oracle provides us with a certificate for the quality of an approximate solution.

- Let $Q \equiv E$. Then for any $g_y \in \partial f(y)$ we have

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

  Indeed, for any $x \in E$ we have $f(x) \geq f(y) + \langle g_y, x - y\rangle \geq f_{\delta,L}(y) + \langle g_y, x - y\rangle$. Comparing this inequality with the second part of (2), we get (6).

- If f_i has (δ_i, L_i)-oracle, $i = 1, 2$, then $f_1 + f_2$ has $(\delta_1 + \delta_2, L_1 + L_2)$-oracle.

In the end of this sections, let us consider two simple examples of inexact oracle. The more serious applications will be given in Section 3.

1. 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 computed at a shifted point y_δ. Let us show that such an oracle can be seen as an (δ, L)-oracle with

$$\delta = M\,\|y - y_\delta\|_E^2, \quad L = 2M.$$

Indeed, the first inequality in (2) is satisfied since for any $x \in Q$ we have

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

Thus, we can take $f_{\delta,L}(y) \overset{\text{def}}{=} f(y_\delta) + \langle \nabla f(y_\delta), y - y_\delta\rangle$, and $g_{\delta,L}(y) \overset{\text{def}}{=} \nabla f(y_\delta)$.

In order to prove the second inequality in (2), note that for all $x \in Q$ we have

$$\begin{aligned} f(x) \;&\overset{(3)}{\leq}\; f(y_\delta) + \langle \nabla f(y_\delta), x - y_\delta\rangle + \tfrac{M}{2}\|x - y_\delta\|_E^2 \\ &=\; f(y_\delta) + \langle \nabla f(y_\delta), y - y_\delta\rangle + \langle \nabla f(y_\delta), x - y\rangle \\ &\quad + \tfrac{M}{2}\|x - y\|_E^2 + \tfrac{M}{2}\|x - y_\delta\|_E^2 - \tfrac{M}{2}\|x - y\|_E^2. \end{aligned}$$

Since $\|\cdot\|_E^2$ is a convex function, $\|x-y_\delta\|_E^2 \leq 2\|y-y_\delta\|_E^2 + 2\|x-y\|_E^2$. Therefore,

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

Thus, we can take $L = 2M$ and $\delta = M\|y-y_\delta\|_E^2$.

2. 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 lower level of smoothness. Let function f be convex and subdifferentiable on Q. For each $y \in Q$ we fix its unique subgradient $g(y)$ (this has nontrivial sense for nonsmooth functions only). Assume that f satisfies the following smoothness condition:

$$\|g(x)-g(y)\|_E^* \quad \leq \quad L_\nu\,\|x-y\|_E^\nu\,, \quad \forall x,y \in Q, \tag{7}$$

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

$$f(x) \quad \leq \quad f(y) + \langle g(y), x-y\rangle + \tfrac{L_\nu}{1+\nu}\,\|x-y\|_E^{1+\nu}\,, \quad \forall x,y \in Q. \tag{8}$$

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

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

$$f(x) - f(y) - \langle g(y), x-y\rangle \quad \leq \quad \tfrac{A(\delta,\nu)}{2}\,\|x-y\|_E^2 + \delta, \quad \forall x,y \in Q. \tag{9}$$

Then, we can apply to these functions the usual first-order methods working with inexact $(\delta, A(\delta,\nu))$-oracle. Comparing (8) and (9), we come to the following definition:

$$A(\delta,\nu) \quad = \quad 2\max_{t\geq 0}\left\{\tfrac{L_\nu}{1+\nu}t^{-1+\nu} - \delta t^{-2}\right\} \overset{(\tau=1/t^2)}{=} 2\max_{\tau>0}\left\{\tfrac{L_\nu}{1+\nu}\,\tau^{\frac{1-\nu}{2}} - \delta\tau\right\}.$$

The optimal value of τ in the later maximization problem is $\tau_* = \left[\frac{L_\nu}{2\delta}\cdot\frac{1-\nu}{1+\nu}\right]^{\frac{2}{1+\nu}}$. Thus,

$$A(\delta,\nu) \quad = \quad 2\tau_*^{\frac{1-\nu}{2}}\left[\tfrac{L_\nu}{1+\nu} - \delta\tau_*^{\frac{1+\nu}{2}}\right] \quad = \quad L_\nu\tau_*^{\frac{1-\nu}{2}} \quad = \quad L_\nu\left[\tfrac{L_\nu}{2\delta}\cdot\tfrac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}. \tag{10}$$

In particular, for $\nu = 0$ (functions with bounded variation of subgradients),

$$A(\delta,0) \quad = \quad \tfrac{L_0^2}{2\delta}. \tag{11}$$

Thus, the *exact* first-order oracle for nonsmooth convex functions can be seen as an $(\delta, \frac{L_0^2}{2\delta})$-oracle. The similar statement is true for functions with Hôlder-continuous gradient (7). Therefore, we can cover the problems with weaker level of smoothness by our analysis of the methods working with (δ, L)-oracle. Note that in this case, δ does not really represent an accuracy of the oracle. The choice of a smaller δ does not cost more, and the answer of the oracle is the same for any δ. However, the corresponding Lipschitz constant grows as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$. These observations give us a possibility to apply any first-order method of smooth convex optimization to non-smooth or weakly smooth functions (see Section 8).

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

Remark 2 *Another typical approach in order to apply first-order method of $F_L^{1,1}(E)$ to a function with a weaker level of smoothness is to smooth the function using averaging of the first-order informations. Assume that E is endowed with the standard euclidean norm. Consider a convex function $f \in F_M^{1,0}$. Let $\delta > 0, z \in E$, and define:*

$$f_\delta(z) = \frac{1}{V_\delta}\int_{\|y-z\|_2\le\delta} f(y)dy$$

$$g_\delta(z) = \nabla f_\delta(z) = \frac{1}{V_\delta}\int_{\|y-z\|_2\le\delta} g(y)dy$$

where V_δ denotes the volume of the Euclidean ball $B_2(z,\delta)$, and $\{g(y) : y \in B(z,\delta)\}$ is a measurable selection of subgradients of f in this ball. As f is convex and Lipschitz-continuous with constant M we have:

$$f(x) \ge f(y) + \langle g(y), x-z\rangle + \langle g(y), z-y\rangle \quad \forall x,y,z \in E$$

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

Averaging now with respect to y these two inequalities, we obtain:

$$f(x) \ge f_\delta(z) + \langle g_\delta(z), x-z\rangle - \delta M \quad \forall x,z \in Z$$

$$f(x) \le f_\delta(z) + \langle g_\delta(z), x-z\rangle + \delta M + \frac{M}{V_\delta}\int_{\|y-z\|_2\le\delta} \|x-y\|_2\, dy.$$

Furthermore, we have:

$$\|x-y\|_2 \le \sqrt{2\|x-z\|_2^2 + 2\|y-z\|_2^2} \le \frac{2\|x-z\|_2^2 + 2\|y-z\|_2^2}{2\delta} + \frac{\delta}{2}.$$

and therefore:

$$f(x) \le f_\delta(z) + \langle g_\delta(z), x-z\rangle + \frac{M\|x-z\|^2}{\delta} + \frac{5M\delta}{2}.$$

With $\delta = \frac{7M}{2}\delta$, $f_{\delta,L}(z) = f_\delta(z) - \delta M$, $g_{\delta,L}(z) = g_\delta(z)$, we obtain an $(\delta, \frac{7M^2}{2\delta})$-oracle. Note that the dependence of L in M and δ is of the same order as what we have using directly subgradients instead of averaging.

3 Inexact oracles obtained by optimization procedures

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

$$f(x) \;=\; \max_{u\in U} \Psi(x,u), \tag{12}$$

where U is a convex set, and $\Psi(x,u)$ is smooth and strongly concave in u for any $x \in Q$ with concavity parameter $\kappa \geq 0$. The computation of $f(x)$ and $\nabla f(x)$ requires the exact solution of this auxiliary problem. However, very often this is impossible or too costly. Instead, we have to use the approximate solutions.

We will measure the accuracy of an approximate solution u_x for problem (12) in three different ways:

$$\begin{array}{rcl} V_1(u_x) & = & \max\limits_{u \in U} \langle \nabla_2 \Psi(x,u_x), u - u_x \rangle, \\[2ex] V_2(u_x) & = & \max\limits_{u \in U} \left[\Psi(x,u) - \Psi(x,u_x) + \frac{\kappa}{2} \| u_x - u \|_E^2 \right], \\[2ex] V_3(u_x) & = & \max\limits_{u \in U} \left[\Psi(x,u) - \Psi(x,u_x) \right]. \end{array} \qquad (13)$$

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

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

Therefore,

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

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

We describe now three classes of max-type functions for which the approximate solution of subproblem (12) must satisfy one of conditions $V_i(u_x) \leq \delta$. The choice of i depends on the class, taking into account the definition of (δ, L)-oracle.

Let us show how to satisfy stopping criterions (13) in practice. The most common criterion is the third one. It reduces to estimating the optimality gap in the value of objective function. In many optimization methods there exists a direct control of this condition. Other criterions are more difficult. Therefore, let us describe a "brut force" approach for satisfying the strongest condition.

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

$$\begin{array}{rcl} 0 & \geq & \langle \nabla_2 \bar{\Psi}(x,u_x^*), u - u_x^* \rangle \;=\; \langle \nabla_2 \bar{\Psi}(x,u_x^*), u_x - u_x^* \rangle + \langle \nabla_2 \bar{\Psi}(x,u_x^*), u - u_x \rangle \\[2ex] & \geq & -\bar{V}_3(u_x) + \langle \nabla_2 \bar{\Psi}(x,u_x^*) - \nabla_2 \bar{\Psi}(x,u_x), u - u_x \rangle + \langle \nabla_2 \bar{\Psi}(x,u_x), u - u_x \rangle \\[2ex] & \geq & -\bar{V}_3(u_x) - \| \nabla_2 \bar{\Psi}(x,u_x^*) - \nabla_2 \bar{\Psi}(x,u_x) \|_* D_u + \langle \nabla_2 \bar{\Psi}(x,u_x), u - u_x \rangle. \end{array}$$

Hence, if $\nabla_2 \bar{\Psi}(x,\cdot)$ is Lipschitz continuous on U with constant L, then we get

$$V_1(u_x) \;\leq\; \bar{V}_1(u_x) + \mu D_u^2 \;\overset{(3)}{\leq}\; \bar{V}_3(u_x) + D_u [2L\bar{V}_3(u_x)]^{1/2} + \mu D_u^2.$$

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 FGM, then it requires at most $O(\frac{L^{1/2}}{\delta^{1/2}} \ln \frac{1}{\delta})$ iterations (see section 2.2 in [13]).

3.1 Functions obtained by smoothing technique

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 (12) 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 defined as

$$L(f) \;=\; \tfrac{1}{\kappa}\|A\|^2_{F\to E^*}. \tag{14}$$

where $\|A\|_{F\to E^*} = \max\{\|Au\|_{E^*} : \|u\|_F = 1\}$. The importance of this class of functions is justified by the smoothing approach for nonsmooth convex optimization (see [14, 15, 16, 4]).

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

$$V_3(u_y) \;=\; \Psi(y,u_y^*) - \Psi(y,u_y) \;\leq\; \tfrac{\delta}{2}. \tag{15}$$

Let us show that then we can construct an $(\delta, 2L(f))$-oracle.

Indeed, since $\Psi(\cdot,u)$ is convex, for all $u \in U$, we have

$$\begin{aligned} f(x) \;&=\; \Psi(x,u_x^*) \;\geq\; \Psi(x,u_y) \;\geq\; \Psi(y,u_y) + \langle \nabla_1\Psi(y,u_y), x-y\rangle \\ &=\; f_{\delta,L}(y) + \langle g_{\delta,L}(y), x-y\rangle, \end{aligned} \tag{16}$$

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

Further, note that

$$\langle \nabla_1\Psi(y,u_y^*), x-y\rangle \;=\; \langle g_{\delta,L}(y), x-y\rangle + \langle A(u_y^* - u_y), x-y\rangle. \tag{17}$$

Since f has Lipschitz-continuous gradient, we have:

$$\begin{aligned} f(x) \;&\leq\; f(y) + \langle \nabla f(y), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 \\ &=\; f(y) + \langle \nabla\Psi_1(y,u_y^*), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 \\ &\overset{(17)}{=}\; f(y) + \langle g_{\delta,L}(y), x-y\rangle + \tfrac{L(f)}{2}\|x-y\|_E^2 + \langle A(u_y^* - u_y), x-y\rangle. \end{aligned}$$

On the other hand, we have:

$$\begin{aligned}\langle A(u_y^* - u_y), x - y\rangle &\leq \left\|u_y^* - u_y\right\|_F \left\|A^T(x-y)\right\|_E \\ &\overset{(14)}{\leq} \tfrac{\kappa}{2}\left\|u_y^* - u_y\right\|_F^2 + \tfrac{L(f)}{2}\|x-y\|_E^2 .\end{aligned}$$

Therefore,

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

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

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

In view of conditions (15) and (16), we prove that the pair $(\Psi(y, u_y), Au_y)$, satisfying condition (15), corresponds to an (δ, L)-oracle with $L = 2L(f)$.

3.2 Moreau-Yosida regularization

In this section, we consider functions of the form

$$f(x) \quad = \quad \min_{u\in U}\left\{\mathcal{L}(x,u) \overset{\text{def}}{=} h(u) + \tfrac{\kappa}{2}\|u-x\|_2^2\right\}, \tag{18}$$

where h is a smooth convex function on a convex set $U \subset E$. 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 (18). The Lipschitz constant of the gradient is equal to κ.

Instead of solving exactly the problem (18), we compute a feasible solution u_x satisfying

$$V_2(u_x) \quad = \quad \max_{u\in U}\left\{\mathcal{L}(x,u_x) - \mathcal{L}(x,u) + \tfrac{\kappa}{2}\|u-u_x\|_2^2\right\} \;\leq\; \delta. \tag{19}$$

(Since $\mathcal{L}$ is convex in u, we inverted the sign in the definition of V_2 in (13).) Let us show that for all $x \in Q$ the objects

$$\begin{aligned} f_{\delta,L}(x) &= \mathcal{L}(x,u_x) - \delta \;=\; h(u_x) + \tfrac{\kappa}{2}\|u_x - x\|_2^2 - \delta, \\ g_{\delta,L}(x) &= \nabla_1\mathcal{L}(x,u_x) \;=\; \kappa(x-u_x)\end{aligned} \tag{20}$$

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

$$\begin{aligned} f(x) &= \mathcal{L}(x,u_x^*) \;\geq\; \mathcal{L}(y,u_x^*) + \tfrac{\kappa}{2}\langle y-x, 2u_x^* - x - y\rangle \\ &\overset{(19)}{\geq} \mathcal{L}(y,u_y) + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta + \tfrac{\kappa}{2}\langle y-x, 2u_x^* - x - y\rangle \\ &= \mathcal{L}(y,u_y) + \kappa\langle y-u_y, x-y\rangle + \tfrac{\kappa}{2}\|u_x^* - u_y\|_2^2 - \delta \\ &\quad + \tfrac{\kappa}{2}\langle y-x, 2u_x^* - 2u_y + y - x\rangle \\ &= \mathcal{L}(y,u_y) + \kappa\langle y-u_y, x-y\rangle - \delta \\ &\quad + \tfrac{\kappa}{2}\left(\|u_x^* - u_y\|_2^2 + \|y-x\|_2^2 + 2\langle y-x, u_x^* - u_y\rangle\right) \\ &\geq \mathcal{L}(y,u_y) + \kappa\langle y-u_y, x-y\rangle - \delta.\end{aligned}$$

Thus, we satisfy the first inequality in (2) with the values defined by (20).

Further, for all $x, y \in Q$ we have

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

Thus, in view of definition (20), we prove the second inequality in (2) with $L = \kappa$.

3.3 Functions defined by Augmented Lagrangians

Consider the following convex problem:

$$\max_{u\in U}\left\{h(u) :\; Au = 0\right\}, \tag{21}$$

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

$$\min_{x\in E}\; f(x), \tag{22}$$

$$f(x) \;\overset{\text{def}}{=}\; \max_{u\in U}\left[\Psi(x,u) \overset{\text{def}}{=} h(u) + \langle x, Au\rangle - \tfrac{\kappa}{2}\|Au\|_2^2\right]. \tag{23}$$

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 (23). The Lipschitz constant of the gradient is equal to $\frac{1}{\kappa}$.

The problem (22) is usually solved by a first-order method. For that, we need to compute exactly $f(x_k)$ and $\nabla f(x_k)$ at each test point x_k, which is impossible or to costly in practice.

Assume instead, that we compute an approximation $u_x \in U$ such that

$$\begin{aligned} V_1(u_x) &= \max_{u\in U}\,\langle \nabla_2\Psi(x,u_x), u - u_x\rangle \\ &= \max_{u\in U}\,\langle \nabla h(u_x) + A^T x - \kappa A^T A u_x, u - u_x\rangle \;\leq\; \delta. \end{aligned} \tag{24}$$

Let us show that the objects

$$f_{\delta,L}(x) \;=\; \Psi(x,u_x), \quad g_{\delta,L}(x) \;=\; \nabla_1\Psi(x,u_x) \;=\; Au_x \tag{25}$$

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

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

Thus, in view of definition (25), the first inequality in (2) is proved. Further,

$$\begin{array}{rcl} f(x) & \leq & \max\limits_{u\in U}\{h(u_y) + \langle \nabla h(u_y), u-u_y\rangle + \langle x, Au\rangle - \frac{\kappa}{2}\|Au\|_2^2\} \\ & \overset{(24)}{\leq} & \max\limits_{u\in U}\{h(u_y) - \langle A^Ty - \kappa A^TAu_y, u-u_y\rangle + \langle x, Au\rangle - \frac{\kappa}{2}\|Au\|_2^2\} + \delta \\ & = & \Psi(y,u_y) + \langle Au_y, x-y\rangle \\ & & + \max\limits_{u\in U}\left\{\langle x-y, A(u-u_y)\rangle - \frac{\kappa}{2}\|A(u-u_y)\|_2^2\right\} + \delta. \end{array}$$

Thus, in view of (25), we prove the second inequality in (2) with $L = \frac{1}{\kappa}$.

4 Gradient methods with inexact oracle

Consider the problem (1), where f is endowed with (δ, L)-oracles. In this section, we will use the standard Euclidean norm $\|x\|_2 = \langle x, x\rangle^{1/2}$. We assume that the gradient mapping

$$T_L(x,g) \quad = \quad \arg\min_{y\in Q}[\tfrac{1}{L}\langle g, y-x\rangle + \tfrac{1}{2}\|y-x\|_2^2]$$

is computable. The first order optimality condition for point $T_L(x,g)$ are as follows:

$$\langle g + L(T_L(x,g)-x), y - T_L(x,g)\rangle \quad \geq \quad 0 \quad \forall y\in Q. \tag{26}$$

4.1 Primal gradient method (PGM)

Consider the following method:

$$\begin{array}{rl} \textbf{Initialization:} & \text{Choose } x_0\in Q. \\ \textbf{Iteration } (k\geq 0)\textbf{:} & \text{Choose } \delta_k \text{ and } L_k. \\ & \text{Compute } (f_{\delta_k,L_k}(x_k), g_{\delta_k,L_k}(x_k)). \\ & \text{Compute } x_{k+1} = T_{L_k}(x_k, g_{\delta_k,L_k}(x_k)). \end{array} \tag{27}$$

Lemma 1 *For $k\geq 1$, we have*

$$\sum_{i=0}^{k-1}\tfrac{1}{L_i}[f(x_{i+1}) - f(x^*)] \quad \leq \quad \tfrac{1}{2}\|x_0 - x^*\|^2 + \sum_{i=0}^{k-1}\tfrac{\delta_i}{L_i}. \tag{28}$$

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

$$\begin{aligned}
r_{k+1}^2 &= r_k^2 + 2\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle - \|x_{k+1} - x_k\|^2 \\
&\overset{(26)}{\leq} r_k^2 + \tfrac{2}{L_k}\langle g_k, x^* - x_{k+1}\rangle - \|x_{k+1} - x_k\|^2 \\
&= r_k^2 + \tfrac{2}{L_k}\langle g_k, x^* - x_k\rangle - \tfrac{2}{L_k}[\langle g_k, x_{k+1} - x_k\rangle + \tfrac{L_k}{2}\|x_{k+1} - x_k\|^2] \\
&\overset{(2)}{\leq} r_k^2 + \tfrac{2}{L_k}[f(x^*) - f_k] - \tfrac{2}{L_k}[f(x_{k+1}) - f_k - \delta_k].
\end{aligned}$$

Summing up these inequalities for $i = 0, \dots, k-1$, we obtain (28). □

When the exact first-order information is used ($\delta_i = 0$, $L_i = L$), then the sequence $\{f(x_i)\}$ is a decreasing sequence. It is not true when we use an inexact oracle. Therefore, let us define

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

Since f is convex,

$$f(\hat{x}_k) - f(x^*) \;\leq\; \frac{\frac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k-1} L_i^{-1}\delta_i}{\sum_{i=0}^{k-1} L_i^{-1}}. \tag{29}$$

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

$$f(\hat{x}_k) - f(x^*) \;\leq\; \frac{LR^2}{2k} + \delta, \quad R \overset{\mathrm{def}}{=} \|x_0 - x^*\|_2\,. \tag{30}$$

Thus, there is no error accumulation, and the upper bound for the residual is decreasing with k up to the level δ. Hence, for the accuracy of order δ, we need $O(\frac{LR^2}{\delta})$ iterations.

4.2 Dual gradient method [17]

This method 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 $(f_{\delta_k,L_k}(x_k), g_{\delta_k,L_k}(x_k))$. (31)

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

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

Lemma 2 *For any $k \geq 0$ we have*

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

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

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

In view of the first inequality in (2), for all $x \in Q$ we have

$$\psi_k^* \;\leq\; \psi_k(x) \;\leq\; \textstyle\sum_{i=0}^{k} \frac{1}{L_i} f(x) + \frac{1}{2}\|x - x_0\|^2 . \tag{33}$$

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

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

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

$$\psi_k(x) \;\geq\; \psi_k^* + \tfrac{1}{2}\|x - x_k\|_2^2, \quad x \in Q$$

Therefore,

$$\begin{aligned} \psi_{k+1}^* \;&=\; \min_{x \in Q} \left\{ \psi_k(x) + \tfrac{1}{L_k}[f_k + \langle g_k, x - x_k\rangle] \right\} \\ &\geq\; \psi_k^* + \tfrac{1}{L_k} \min_{x \in Q} \left\{ f_k + \langle g_k, x - x_k\rangle + \tfrac{L_k}{2}\|x - x_k\|_2^2 \right\} \\ &\overset{(2)}{\geq}\; \psi_k^* + \tfrac{1}{L_k}(f(y_k) - \delta_k). \end{aligned}$$

Thus, using our inductive assumption, we prove that $\psi_k^* \geq \sum\limits_{i=0}^{k} \frac{1}{L_i}[f(y_i) - \delta_i]$ for all $k \geq 0$. It remains to combine this fact with inequality (33) for $x = x^*$. □

Same as for Primal Gradient Method, we can 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 decreasing upper bound

$$f(\hat{y}_k) - f(x^*) \;\leq\; \frac{\frac{1}{2}\|x_0 - x^*\|_2^2 + \sum_{i=0}^{k} L_i^{-1}\delta_i}{\sum_{i=0}^{k} L_i^{-1}}, \quad k \geq 0. \tag{34}$$

Thus, we obtain the same convergence results as that for PGM. For this reason, in the rest of this paper notation PGM refers both to the primal and to the dual gradient methods.

5 Fast gradient method with inexact oracle

In this section, we adapt one of the last versions of Fast Gradient Method (FGM) developed in [14]. Let $d(x)$ be a prox-function, which is differentiable and strongly convex on Q, and $x_0 = \arg\min\limits_{x\in Q} d(x)$ be its prox-center. By translating and scaling d if necessary, we can always ensure that

$$d(x_0) = 0, \quad d(x) \geq \tfrac{1}{2}\|x - x_0\|^2, \quad \forall x \in Q. \tag{35}$$

Here $\|\cdot\|$ denotes any norm on E.

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

$$\alpha_0 \;\in\; (0,1], \quad \tfrac{\alpha_k^2}{L_k} \;\leq\; A_k \;\overset{\text{def}}{=}\; \sum_{i=0}^{k} \tfrac{\alpha_i}{L_i}, \quad k \geq 0. \tag{36}$$

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

$$\begin{array}{l} \textbf{Initialization:} \quad \text{Choose } \delta_0,\ L_0, \text{ and } x_0 = \arg\min\limits_{x\in Q} d(x). \\[2ex] \textbf{Iteration } (k \geq 0)\textbf{:} \quad \textbf{1.} \text{ Compute } (f_{\delta_k,L_k}(x_k), g_{\delta_k,L_k}(x_k)). \\[2ex] \textbf{2.} \text{ Compute } y_k = T_{L_k}(x_k, g_{\delta_k,L_k}(x_k)). \\[2ex] \textbf{3.} \text{ Compute } z_k = \arg\min\limits_{x\in Q}\{d(x) + \sum\limits_{i=0}^{k} \frac{\alpha_i}{L_i}\langle g_{\delta_i,L_i}(x_i), x - x_i\rangle\}. \\[2ex] \textbf{4.} \text{ Choose } \delta_{k+1} \text{ and } L_{k+1}. \text{ Define } x_{k+1} = \tau_k z_k + (1-\tau_k) y_k. \end{array} \tag{37}$$

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

Theorem 1 *For all $k \geq 0$, we have: $A_k f(y_k) \leq \psi_k^* + E_k$ with $E_k = \sum\limits_{i=0}^{k} A_i \delta_i$.*

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

$$\begin{aligned} \psi_0^* &= \min_{x\in Q}\left\{d(x) + \tfrac{\alpha_0}{L_0}[f_0 + \langle g_0, x - x_0\rangle]\right\} \\ &\overset{(35)}{\geq} \tfrac{\alpha_0}{L_0}\min_{x\in Q}\left\{f_0 + \langle g_0, x - x_0\rangle + \tfrac{L_0}{2}\|x - x_0\|^2\right\} \overset{(2)}{\geq} \tfrac{\alpha_0}{L_0}[f(y_0) - \delta_0]. \end{aligned}$$

Assume now that the statement of the theorem is true for some $k \geq 0$. By the optimality conditions of the optimization problem at Step 3,

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

Hence, in view of strong convexity of d,

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

Thus, for all $x \in Q$,

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

We have obtained:

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

On the other hand, we have:

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

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

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

Therefore,

$$\begin{aligned} \psi_{k+1}^* &\geq A_{k+1} f_{k+1} - E_k + \min_{x \in Q}\{\tfrac{1}{2}\|x - z_k\|^2 + \tfrac{\alpha_{k+1}}{L_{k+1}} \langle g_{k+1}, x - z_k \rangle\} \\ &= A_{k+1}\left[f_{k+1} + \min_{x \in Q}\{\tfrac{1}{2A_{k+1}}\|x - z_k\|^2 + \tau_k \langle g_{k+1}, x - z_k \rangle\}\right] - E_k \\ &\overset{(36)}{\geq} A_{k+1}\left[f_{k+1} + \min_{x \in Q}\{\tfrac{\tau_k^2 L_{k+1}}{2}\|x - z_k\|^2 + \tau_k \langle g_{k+1}, x - z_k \rangle\}\right] - E_k. \end{aligned}$$

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

$$\begin{aligned} & \min_{x\in Q} \left\{ \tfrac{\tau_k^2 L_{k+1}}{2} \|x - z_k\|^2 + \tau_k \langle g_{k+1}, x - z_k\rangle \right\} \\ = \; & \min_{y} \left\{ \tfrac{L_{k+1}}{2} \|y - x_{k+1}\|^2 + \langle g_{k+1}, y - x_{k+1}\rangle : \; y \in \tau_k Q + (1-\tau_k) y_k \right\} \qquad (38) \\ \geq \; & \min_{y\in Q} \left\{ \tfrac{L_{k+1}}{2} \|y - x_{k+1}\|^2 + \langle g_{k+1}, y - x_{k+1}\rangle \right\}. \end{aligned}$$

On the other hand,

$$\begin{aligned} \Psi^*_{k+1} \quad & \geq \quad A_{k+1} \left[f_{k+1} + \min_{x\in Q} \{ \tfrac{\tau_k^2 L_{k+1}}{2} \|x - z_k\|^2 + \tau_k \langle g_{k+1}, x - z_k\rangle \} \right] - E_k \\ & \overset{(2),(38)}{\geq} \quad A_{k+1} f(y_{k+1}) - E_k - A_{k+1}\delta_{k+1}, \end{aligned}$$

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 2 *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

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

Now, using the recurrence obtained in Theorem 1, we complete the proof. □

If use the simplest choice of sequence $\{\alpha_i\}$, i.e $\alpha_i = \frac{i+1}{2}$, then the sequence of Lipschitz constants must satisfy inequality $\frac{(k+1)^2}{4L_k} \overset{(36)}{\leq} \sum\limits_{i=0}^k \frac{i+1}{2L_i}$, i.e.

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

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

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

Consider the case of constant accuracy of the oracle ($\delta_i = \delta$, $L_i = L$). Then we have $A_k = \frac{(k+1)(k+2)}{4L}$, $\tau_k = \frac{2}{k+3}$, and therefore

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

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

$$f(y_k) - f^* \;\leq\; \tfrac{4Ld(x^*)}{(k+1)(k+2)} + \tfrac{1}{6}(2k+6)\delta \;\leq\; \tfrac{4LR^2}{(k+1)^2} + \tfrac{1}{3}(k+3)\delta. \tag{39}$$

Contrarily to the classical gradient methods, the use of inexact oracle in FGM results in accumulation of errors. The first terms in (39) decreases as $O(\frac{1}{k^2})$, but the second term is increasing in k. Asymptotically, the use of inexact oracle makes FGM divergent.

For non-asymptotic behavior, we can consider two cases.

1. The oracle accuracy δ is fixed.

In this case, we can find the number of iterations k^* that minimizes the residual in the objective function:

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

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

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

At this moment, the obtained accuracy in the objective function is:

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

2. The oracle accuracy δ can be chosen.

Let us assume that parameter L of inexact oracle is independent on δ. If we need to reach the 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 $k \leq \frac{3\epsilon}{2\delta} - 3$. This is possible if and only if

$$\delta \;\leq\; \tfrac{3\epsilon^{3/2}}{2\sqrt{8Ld(x^*)}+4\sqrt{\epsilon}}. \tag{40}$$

In conclusion, if we choose the oracle accuracy satisfying relation (40), then after

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

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

Contrarily to the classical gradient methods, in order to reach accuracy ϵ by FGM, we need to require that the accuracy of inexact oracle satisfies (40).

6 Inexact oracle: What method is better?

If the oracle is exact, FGM is the optimal method for the class $F_L^{1,1}(Q)$. For the accuracy ϵ in the objective function, it needs $O(\sqrt{\frac{L}{\epsilon}}R)$ iterations. At the same time, PGM needs $O\left(\frac{LR^2}{\epsilon}\right)$ iterations.

Situation is more complicated when inexact first-order oracle is used. Contrary to PGM, FGM suffers from an errors accumulation. In order to compare their efficiency, we consider two cases.

1. Oracle accuracy can be chosen.

In this case we assume that L is independent on the oracle accuracy δ (see examples in Section 3). If we need to reach the accuracy of ϵ for function value, PGM with inexact oracle requires accuracy of the oracle on the level of $\Theta(\epsilon)$. However, it needs $O\left(\frac{LR^2}{\epsilon}\right)$ iterations.

For FGM with inexact oracle, due to the errors accumulation, we need a higher accuracy of the oracle ($\Theta\left(\frac{\epsilon^{3/2}}{\sqrt{L}R}\right)$). But the necessary number of iterations is only of the order $O\left(\sqrt{\frac{L}{\epsilon}}R\right)$. Thus, the choice between two methods depends on complexity of inexact oracle. Denote by $C(\delta)$, the computational time, which is needed by inexact oracle for computing the answer $(f_{\delta,L}(x), g_{\delta,L}(x))$. Then PGM is preferable if

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

Consider the following situations.

- The oracle is very expensive: $C(\delta) = \Omega\left(\frac{1}{\delta}\right)$ (e.g. $C(\delta) = \frac{1}{\delta^2}$). Then, it is preferable to use PGM.
- Oracle has moderate efficiency: $C(\delta) = \Theta\left(\frac{1}{\delta}\right)$. Then both methods are in a certain sense equivalent.
- Oracle is very efficient: $C(\delta) = o\left(\frac{1}{\delta}\right)$ (for example, $C(\delta) = \frac{1}{\delta^{1/2}}$, or even $C(\delta) = \ln\frac{1}{\delta}$). Then FGM is better.

2. Oracle accuracy is fixed.

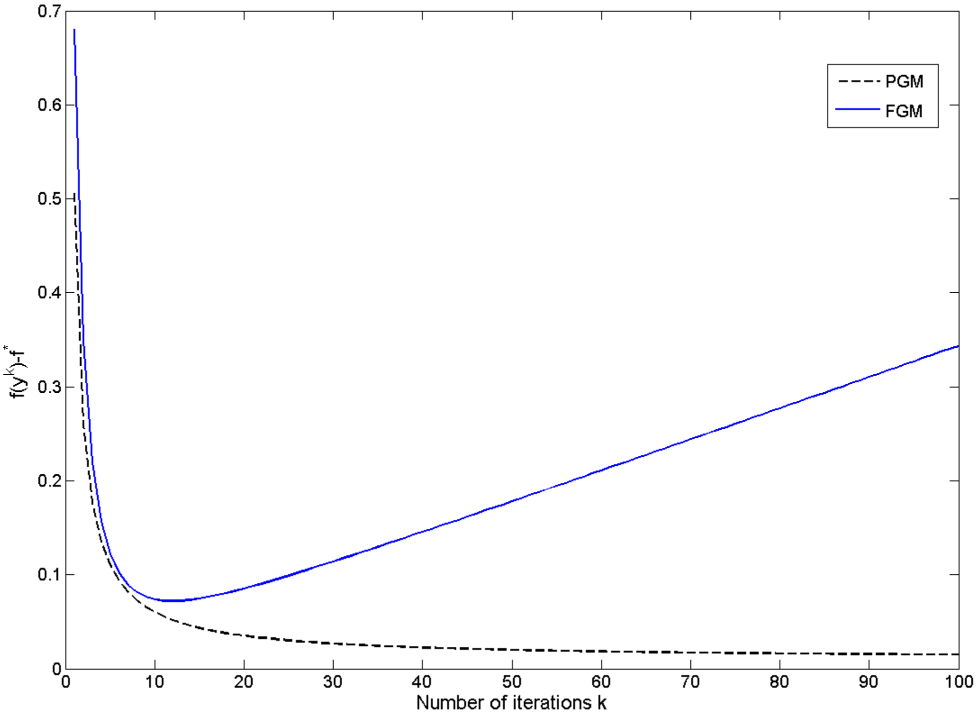
In this case, the sequence of iterates generated by PGM satisfies inequality

$$f(x^k) - f^* \quad \le \quad \tfrac{LR^2}{2k} + \delta,$$

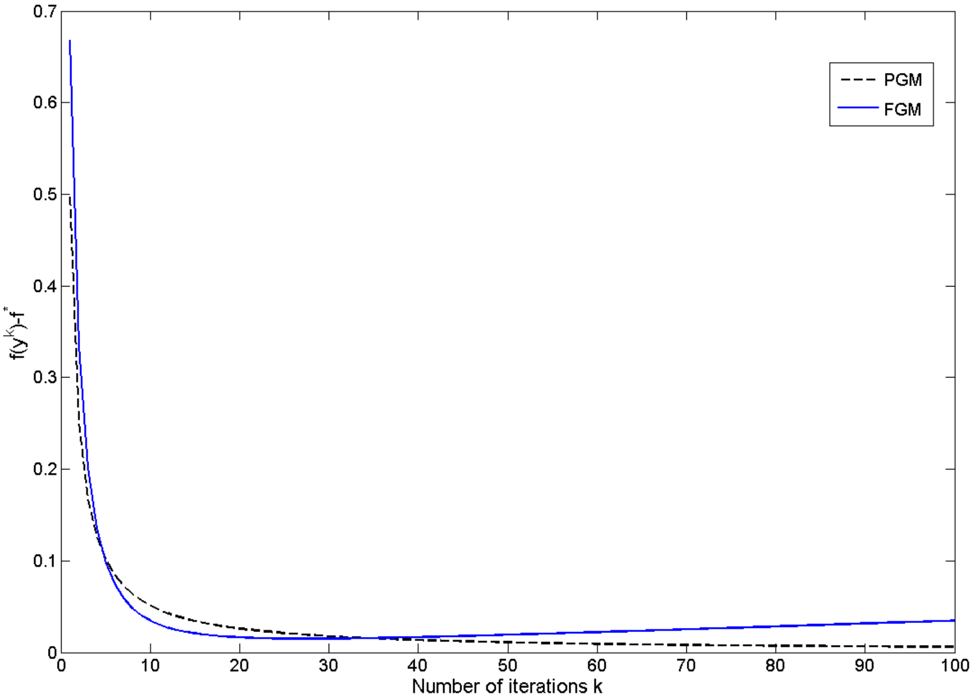
whereas the sequence obtained by FGM satisfies inequality

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

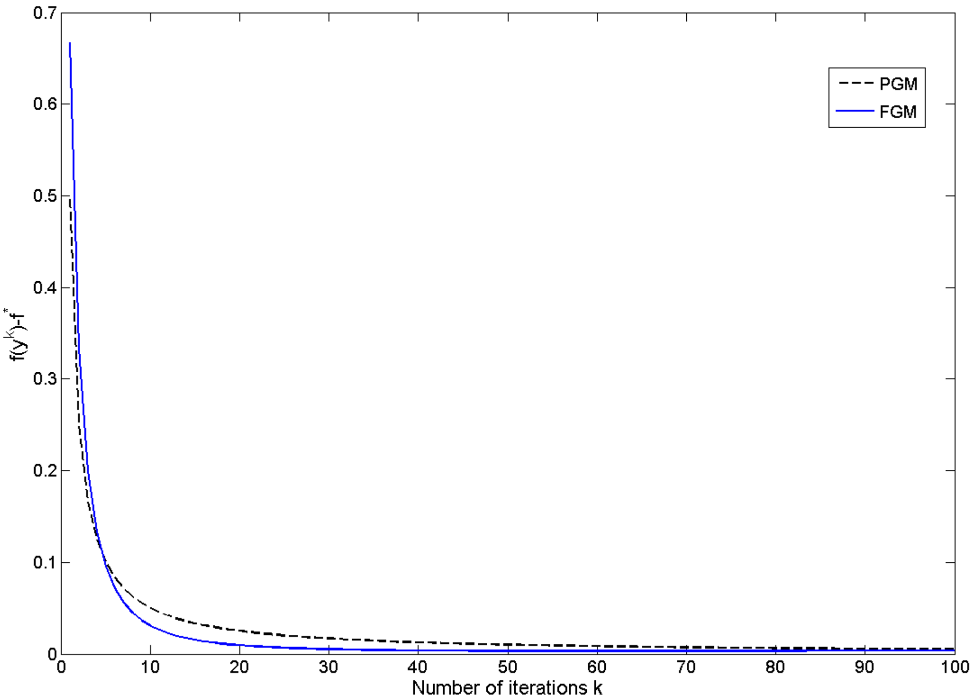
The dependence of these two rates of convergence in k are represented at the following picture for different values of the oracle accuracy:



$\delta = 0.01$, $L = 1$ and $R = 1$



$\delta = 0.001$, $L = 1$ and $R = 1$



$\delta = 0.0001$, $L = 1$ and $R = 1$

The higher is the accuracy of the oracle, the larger is the number of iterations, where the FGM is better than PGM. At the limit, when the oracle accuracy $\delta = 0$, FGM outperforms PGM for any number of iterations.

On the other hand, when the oracle accuracy is very poor, the accumulation of oracle errors in the FGM is so high that PGM is always better than FGM.

For a higher, but non zero accuracy, the situation is more complicated. Due to the constant factors in the convergence rates, PGM starts to provide smaller error for the few first iterations. After that, due to its high convergence rate, FGM decreases the gap in objective function much better than PGM. For FGM, this gap attains its minimum value at the moment $N_1 = \Theta\left(\sqrt[3]{\frac{LR^2}{\delta}}\right)$ with corresponding accuracy $\delta^* = \Theta(\delta^{2/3}L^{1/3}R^{1/3})$. It is not interesting to perform more iterations since then the gap can only increase due to accumulation of errors.

Note that there exists a moment N_2, after which the PGM provides us better accuracy than FGM. However we have to wait until $N_3 = \Theta\left(\frac{LR^2}{\delta^{2/3}}\right)$ iterations in order to have the accuracy of PGM becomes better than δ^*. After that, we can reach by PGM the final accuracy of order δ (not possible by FGM). This needs $\Theta\left(\frac{LR^2}{\delta}\right)$ iterations. Asymptotically, by PGM we can obtain an accuracy exactly equal to δ.

In conclusion, FGM is the method of choice when we need an accuracy lower than $\delta^{2/3}L^{1/3}R^{2/3}$. For reaching a better accuracy, PGM must be used.

7 Comparison with other approaches

Fast-gradient methods using inexact first-order oracle have been already studied in [3] and [1]. In this approach, the set Q is assumed to be bounded, and oracle provides at each point $y \in Q$, an approximative gradient $g(y)$ satisfying condition

$$|\langle g(y) - \nabla f(y), x - z\rangle| \quad \leq \quad \xi \quad \forall x, y, z \in Q. \tag{41}$$

Let us compare this definition with (2), taking into account their applicability and the obtained results.

First of all, the applicability of (41) needs more assumptions.

- The set Q must be bounded (do not need this for (2)).
- The objective function must be differentiable. The existence of the gradient at all points is necessary since it must be compared with the approximative gradient. In our case, we are able to consider also non-smooth convex function.

But even in the smooth case $f \in F_L^{1,1}(Q)$, we can argue that the condition (41) is stronger than (2). Let $f \in F_L^{1,1}(Q)$:

1. Let us show that the approximative gradient $g(y)$ satisfying (41), can be used also in our definition. Indeed, in view of (3) and (41), for all $x, y \in Q$ we have

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

   Thus, taking $f_{\delta,L}(y) = f(y) - \xi$, and $g_{\delta,L}(y) = g(y)$ we satisfy (2) with $\delta = 2\xi$.

2. On the other hand, our condition (2) does not imply (41) with any $\xi = \Theta(\delta)$. Indeed, consider the function $f(x) = \max_{u\in U} \Psi(x,u)$, where

$$\Psi(x,u) = -\tfrac{1}{2}\|u\|_2^2 + \langle x, u\rangle, \quad Q = \{y \in \mathbb{R}^n : \|y\|_2 \leq 1\}, \quad U = \mathbb{R}^n. \tag{42}$$

   For point $x = 0$, let us fix for the answer of oracle some point u_0 with $\|u_0\|_2 = \delta^{1/2}$. Since $u_0^* = 0$, and $f(0) - \Psi(0, u_{x_0}) = \frac{1}{2}\|u_{x_0}\|_2^2 = \frac{\delta}{2}$, the answer of the oracle $(f_{\delta,L}(0), g_{\delta,L}(0)) = (-\frac{\delta}{2}, u_0)$ satisfies condition (2) with $L = 2$ (see Section 3.1). However,

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

Let us compare now the quality of the answers of these oracles for FGM (we assume that Q is bounded). It is proved that FGM using the oracle (41) converges as follows:

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

where C is an absolute constant. Thus, there is no error accumulation. Therefore, the accuracy of the oracle can be of the same order as the desired accuracy of the solution. At first sight, this result seems to be better than the results obtained with (δ, L)-oracle.

However, note that for the same level of accuracy, condition (41) is much stronger than (2). Let us look at important example. Consider the class of functions with explicit max-structure: $f(x) = \max_{u\in U} \Psi(x,u)$, where set U is closed and convex, and $\Psi(x,u) =$

$G(u) + \langle x, Au\rangle$, where $G(u)$ is a differentiable, strongly concave function with concavity parameter κ. Assume that we want to solve the primal problem $\min_{x\in Q} f(x)$ with accuracy ϵ. In our definition of inexact oracle, the oracle accuracy δ corresponds directly to the accuracy of solving the dual problem (see Section 3.1).

For definition (41), we can also use an approximate dual solution:

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

However, now we need to satisfy the following relation:

$$|\langle A(u_x^* - u_x), y - z\rangle| \;\; \leq \;\; \epsilon, \quad \forall x, y, z \in Q. \tag{43}$$

(We can take $\xi = \epsilon$ since the condition (41) avoids accumulation of errors). For that, we need to have u_x close to u_x^*:

$$\|u_x - \overline{u}_x\| \;\; \leq \tfrac{\epsilon}{\text{diam}\,(Q)\cdot\|A\|_{F\to E^*}}.$$

Since Ψ is strongly concave: $\Psi(x, u_x^*) - \Psi(x, u_x) \geq \frac{\kappa}{2}\,\|u_x - u_x^*\|^2$, a sufficient condition for (41) is as follows:

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

In our approach, in order to avoid accumulation of errors, it is enough to solve the dual problem up to accuracy $\epsilon^{3/2}$ (see (40)) (instead of ϵ^2 for 41) .

Remark 3 *In some cases, inequality $\Psi(x, u_x^*) - \Psi(x, u_x) \leq \epsilon^2/8$ is also a necessary condition for (43). Indeed, consider again the saddle point problem defined by (42). We have $f(0) - \Psi(0, u_0) = \frac{1}{2}\,\|u_0\|_2^2$. In order to satisfy condition (43) we need to ensure*

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

Remark 4 *The definition of inexact oracle used in [1] is in fact a little bit different from (41). The author assumes that $g(y)$ satisfies the following conditions:*

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

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

and that the set Q is bounded. It is possible to prove that this definition implies (41) with $\xi = D_Q\overline{\xi}$ (where D_Q denotes the diameter of Q) and with $\nabla f(y)$ eventually replaced by a subgradient when the function is non-smooth.

8 Applications to non-smooth optimization

8.1 Solving weakly smooth problems

Let f be a convex function satisfying the Hôlder condition (7). This class includes non-smooth convex functions with bounded variation of subgradients ($\nu = 0$), and smooth convex functions with Hôlder 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 the Lipschitz constants that grow as $O\left(\delta^{-\frac{1-\nu}{1+\nu}}\right)$ with the desired oracle accuracy.

Remark 5 *In this case, δ does not represent the real accuracy of the oracle. It does not cost more to generate a first-order information corresponding to a smaller δ. In fact, for each $\delta > 0$, the answer of (δ, L)-oracle is the same. It just returns the value of the function and a subgradient.*

Oracle accuracy δ is involved only in the computation of Lipschitz constant $L = A(\delta,\nu)$. This constant must be properly used in the numerical methods. In view of this flexibility, there is always a tradeoff between the high "accuracy" of the oracle, and the small Lipschitz constant L.

For the sake of simplicity, we assume that the number of iterations N is fixed.

Let us apply PGM (27) to a weakly smooth function f with the inexact (δ, L)-oracle. In view of (30), after N iterations we have

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

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:

$$f(\hat{x}_N) - f(x^*) \;\;\leq\;\; \delta_N \left(\tfrac{C_N}{\delta_N^{1+\tau}} + 1\right) \;\;=\;\; \tfrac{2\delta_N}{1-\nu}. \tag{44}$$

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:

$$f(\hat{x}_N) - f(x^*) \;\;\leq\;\; \tfrac{L_\nu R^{1+\nu}}{1+\nu} \cdot \left(\tfrac{2}{N}\right)^{\frac{1+\nu}{2}} . \tag{45}$$

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

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

which is the optimal rate of convergence (see [12, 13]). However for functions with Hôlder continuous gradient, the obtained rate is not optimal (it can reach $O(N^{-\frac{1+3\nu}{2}})$, see [11, 9]).

Further, let us apply now FGM to a weakly smooth function using an (δ, L)-oracle. In view of (39), after N iterations we have:

$$\begin{aligned} f(y_N) - f(x^*) \quad &\leq \quad 4L_\nu \left[\tfrac{L_\nu}{2\delta} \cdot \tfrac{1-\nu}{1+\nu} \right]^{\frac{1-\nu}{1+\nu}} \tfrac{R^2}{(N+1)^2} + \delta \cdot (N+1) \\ &\overset{\rm def}{=} \quad \hat{C}_N \left(\tfrac{1}{\delta} \right)^{\frac{1-\nu}{1+\nu}} + \delta \cdot (N+1). \end{aligned}$$

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

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

Note that

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

Thus, we obtain the following upper bound:

$$f(y_N) - f(x^*) \quad \leq \quad \tfrac{2L_\nu R^{1+\nu}}{1+\nu} \left(\tfrac{2}{N+1} \right)^{\frac{1+3\nu}{2}}. \tag{46}$$

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

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

which is the optimal rate. For functions with Hôlder continuous gradient, the obtained rate of convergence is also optimal ([11, 9]). Thus, we get a universal optimal first-order method both for smooth, weakly smooth and non-smooth convex functions.

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

- We can apply PGM and FGM to objective functions formed as 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 accumulation of errors is an intrinsic property of any FGM. Slower first-order methods can avoid accumulation of errors, and PGM is the fastest method having this good property.

We discuss these topics in the next section.

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 which variation of subgradients 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 (11) function f_2 has $(\delta, \frac{1}{2\delta}M^2(f_2))$-oracle. Hence, we conclude that the data

$$(f_1(y) + f_2(y), \nabla f_1(y) + g_2(y)), \quad g_2(y) \in \partial f_2(y), \tag{47}$$

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

Let us apply now PGM to function f using the inexact (δ, L)-oracle (47). Then, after N iterations we have:

$$f(\hat{x}_N) - f^* \overset{(30)}{\leq} \left(L(f_1) + \tfrac{1}{2\delta}M^2(f_2)\right) \tfrac{R^2}{2N} + \delta.$$

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 nonsmooth part of the problem, but not for the smooth one.

Let us check now the performance of FGM as applied to the composite problems. In view of (39) after N iterations of the scheme, we have

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

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

Remark 6 *Our analysis is similar, in a certain sense, to that of [6], 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 for the non-smooth part of objective, the gradients by subgradients, and the Lipschitz constant by a value of the order $O(M(f_2)N^{3/2})$. 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 accumulation of errors

Applicability of first-order methods of smooth optimization to non-smooth problems, based on the notion of inexact oracle, opens a possibility for deriving lower bounds on accumulation of errors. This is the main subject of this section.

Let us start from the following observation.

Theorem 3 *Consider a first-order method for $F_L^{1,1}(Q)$ with convergence rate $O(\frac{LR^2}{k^p})$. Assume that the bounds on the performance of this method, as applied to a problem equipped with inexact (δ, L)-oracle, are given by inequality*

$$f(z_k) - f^* \quad \leq \quad \tfrac{C_1 L \|x_0 - x^*\|^2}{k^p} + C_2 k^q \delta, \tag{48}$$

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

Proof:
Let f be a non-smooth convex function, which variation of subgradients is 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:

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

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

$$f(z_k) - f^* \quad \leq \quad [2C_1C_2]^{1/2} MR \cdot k^{-\frac{p-q}{2}}.$$

From the lower complexity bounds for nonsmooth optimization problems, we know that the 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, for 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. PGM). In the case of inexact (δ, L)-oracle, the situation is more complicated.

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

From Theorem 3, 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 inexact situation, both FGM and PGM 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 PGM, 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 4 *Let parameter L of inexact oracle (2) be independent on δ. Under assumptions of Theorem 3, 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 (48), 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 1 *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 2 *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 $\delta = \Omega(\frac{\epsilon^{3/2}}{L^{1/2}R})$.

9 Strongly convex case

In this section, we assume that convex function f, which is endowed with (δ,L)-oracle, satisfies also condition

$$f(x) \;\geq\; f(x^*) + \frac{\mu}{2}\|x-x^*\|^2, \quad \forall x \in Q. \tag{49}$$

This inequality is satisfied, for example, when f is strongly convex on Q with parameter μ, that is

$$f(\alpha x + (1-\alpha)y) \;\leq \alpha f(y) + (1-\alpha)f(x) - \alpha(1-\alpha)\frac{\mu}{2}\|x-y\|_E^2 \tag{50}$$

for all $x,y \in Q$, and $\alpha \in [0,1]$. In this section, we study the possibilities of solving the problem (1) with strongly convex objective, when only an inexact (δ,L)-oracle (2) is available.

9.1 Inexact PGM for strongly convex case

Let function f be equipped with (δ, L)-oracle. Let us apply to it PGM, starting from some point $\bar{u} \in Q$. Denote by u_+ the point obtained after N iterations. Then

$$f(u_+) - f^* \overset{(30)}{\leq} \tfrac{L}{2N}\|x^* - \bar{u}\|^2 + \delta \overset{(49)}{\leq} \tfrac{\gamma_f}{N}(f(\bar{u}) - f^*) + \delta,$$

where $\gamma_f = \frac{L}{\mu}$. Choosing $N = 2\gamma_f$, and denoting the resulting u_+ by $p(\bar{u})$, we get

$$f(p(\bar{u})) - f^* - 2\delta \leq \tfrac{1}{2}(f(u_{k-1}) - f^* - 2\delta)$$

Repeating now the operation $u_{k+1} = p(u_k)$, $k \geq 0$, we obtain

$$f(u_k) - f^* \leq \tfrac{1}{2^k}(f(u_0) - f^*) + 2\delta. \tag{51}$$

As usual, in our analysis we consider two cases.

1. The oracle accuracy is fixed.

As in the general convex case, there is no accumulation of errors. The best accuracy for the objective, that can be reached asymptotically, is 2δ. However, we can get the same order of accuracy after iterations in $\Theta(\gamma_f \log_2 \frac{f(x_0)-f^*}{\delta})$ iterations.

2. The oracle accuracy can be chosen

If we need to reach final accuracy ϵ for the residual in objective function, the oracle accuracy can be chosen as $\delta = \frac{1}{4}\epsilon$. Then the process (51) generates the required solution after $k \geq 1 + \log_2 \frac{f(x_0)-f^*}{\epsilon}$ iterations. The total number of iterations in the process does not exceed

$$N \cdot k = 2\gamma_f \cdot \left(1 + \log_2 \tfrac{f(x_0)-f^*}{\epsilon}\right). \tag{52}$$

Note that the dependence of this bound in the condition number γ_f is not optimal.

9.2 Inexact FGM for strongly convex case

Let us apply to problem (1), (2) the fast gradient method starting from some point $\bar{u} \in Q$ and using the prox-function $d_{\bar{u}}(x) = \frac{1}{2}\|x - \bar{u}\|_2^2$. And let u_+ be the point obtained after N iterations. In accordance to (39), we have

$$f(u_+) - f^* \overset{(49)}{=} \tfrac{4\gamma_f(f(\bar{u})-f^*)}{(N+1)^2} + \delta \cdot (N+1).$$

Let us choose $N = 4\gamma_f^{1/2} - 1$ and denote the point u_+ by $v(\bar{u})$. Then,

$$f(v(\bar{u})) - f^* \leq \tfrac{1}{4}(f(u_k) - f^*) + \delta, \quad \bar{\delta} \overset{\text{def}}{=} 4\gamma_f^{1/2} \cdot \delta.$$

Therefore the process $u_{k+1} = v(u_k)$, $k \geq 0$, has the the following convergence:

$$f(u_{k+1}) - f^* - \tfrac{4\bar{\delta}}{3} \leq \tfrac{1}{4}(f(u_k) - f^* - \tfrac{4\bar{\delta}}{3}) \leq \tfrac{1}{4^{k+1}}(f(u_0) - f^* - \tfrac{4\bar{\delta}}{3}). \tag{53}$$

We consider now two important cases.

1. The oracle accuracy is fixed.

Contrarily to the general case, now there is no accumulation of oracle errors. The error on the objective function decreases with the number of iterations. However, we are not able to reach the level of oracle accuracy. The best what can be achieved asymptotically is $\Theta(\gamma_f^{1/2}\delta)$. This needs $\Theta(\gamma_f^{1/2}\log_2\frac{f(u_0)-f^*}{\delta})$ iterations. Thus, FGM is the method of choice when the target accuracy for objective is not higher than $\gamma_f^{1/2}\delta$.

2. The oracle accuracy can be chosen.

If we need accuracy ϵ for the objective, we can define $\delta \leq \frac{9\epsilon}{32\gamma_f^{1/2}}$. Formally, we can choose the oracle accuracy δ of the same order as ϵ. However note that in some applications the value γ_f can be very big. Under this choice, we reach the level ϵ in $k = O(\log_4\frac{f(u_0)-f^*}{\epsilon})$ iterations. Hence, the total number of iterations of FGM does not exceed

$$O(\gamma_f^{1/2}\log_4\tfrac{f(u_0)-f^*}{\epsilon}). \tag{54}$$

Thus, in the strongly convex case, for the choice between PGM and FGM, we have compare the efficiency of the method with the accuracy of the oracle. The picture becomes more diverse since we need to take into account the magnitude of the condition number.

9.3 Application to non-smooth strongly convex problems

In this section we assume that function f, which variation of subgradients is bounded by constant M, is strongly convex with parameter μ. Assume also that its inexact $(\delta, \frac{M^2}{2\delta})$-oracle is available.

In order to solve problem (1) up to accuracy ϵ, let us apply PGM described in Section 9.1. We need to choose $\delta = \frac{1}{4}\epsilon$. Then the condition number is as follows:

$$\gamma_f \;=\; \tfrac{1}{\mu}\cdot\tfrac{2M^2}{\epsilon}.$$

Thus, in accordance to (52), we need $O(\frac{2M^2}{\mu\epsilon}\ln\frac{f(u_0)-f^*}{\epsilon})$ iterations. This complexity is optimal, up to a logarithmic factor (see [12, 5]).

For the fast gradient method, described in Section 9.2, we need to satisfy the system of equations

$$\delta \;=\; \tfrac{9\epsilon}{32\gamma_f^{1/2}}, \quad \gamma_f \;=\; \tfrac{M^2}{2\delta\mu}.$$

Thus, $\gamma_f = \left(\frac{16M^2}{9\mu\epsilon}\right)^2$, and we obtain from (54) the same optimal complexity (up to a logarithmic factor).

These results confirm that our complexity analysis presented in Sections 9.1, 9.2, is tight both for PGM and FGM.

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 approximative 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] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization. *Preprint Florida State University*, (2010)

[6] G. Lan. An optimal method for stochastic composite optimization. *Preprint, submitted to Mathematical Programming* (2009)

[7] M. Hintermuller. A proximal bundle method based on approximative subgradient. *Computational Optimization and Applications*, **20**, 245-266 (2001)

[8] K. Kiwiel. A proximal bundel method with approximative subgradient linearization. *SIAM Journal of Optimization*, **16**, 1007-1023 (2006)

[9] 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).

[10] A. Nedic and D.Bertsekas. The effect of deterministic noise in subgradient methods. *Mathematical programming, Serie A*, **125**, 75-99 (2010).

[11] A. Nemirovskii and Y.Nesterov. Optimal methods for smooth convex minimization. *Zh. Vichisl. Mat. Fiz. (In Russian)*, **25(3)**, 356-369 (1985).

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

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

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

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

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

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

[18] 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).

[19] Yu. Nesterov.On an approach to the construction of optimal methods of minimization of smooth convex function, *È konom. i. Mat. Metody (In Russian)*, **24**, 509-517 (1988).

[20] B.T. Polyak. Introduction to Optimization. *Optimization Software Inc* (1987)

[21] N.Z. Shor. Minimization Methods for Non-Differentiable Functions. *Springer Series in Computational Mathematics. Springer-Verlag* (1985).

[22] P. Tseng. On accelerated Proximal Gradient Methods for Convex-Concave Optimization *Submitted to Siam. J. Optim.* (2008).

Recent titles

CORE Discussion Papers

2010/50. Maik SCHWARZ, Sébastien VAN BELLEGEM and Jean-Pierre FLORENS. Nonparametric frontier estimation from noisy data.

2010/51. Nicolas GILLIS and François GLINEUR. On the geometric interpretation of the nonnegative rank.

2010/52. Yves SMEERS, Giorgia OGGIONI, Elisabetta ALLEVI and Siegfried SCHAIBLE. Generalized Nash Equilibrium and market coupling in the European power system.

2010/53. Giorgia OGGIONI and Yves SMEERS. Market coupling and the organization of counter-trading: separating energy and transmission again?

2010/54. Helmuth CREMER, Firouz GAHVARI and Pierre PESTIEAU. Fertility, human capital accumulation, and the pension system.

2010/55. Jan JOHANNES, Sébastien VAN BELLEGEM and Anne VANHEMS. Iterative regularization in nonparametric instrumental regression.

2010/56. Thierry BRECHET, Pierre-André JOUVET and Gilles ROTILLON. Tradable pollution permits in dynamic general equilibrium: can optimality and acceptability be reconciled?

2010/57. Thomas BAUDIN. The optimal trade-off between quality and quantity with uncertain child survival.

2010/58. Thomas BAUDIN. Family policies: what does the standard endogenous fertility model tell us?

2010/59. Nicolas GILLIS and François GLINEUR. Nonnegative factorization and the maximum edge biclique problem.

2010/60. Paul BELLEFLAMME and Martin PEITZ. Digital piracy: theory.

2010/61. Axel GAUTIER and Xavier WAUTHY. Competitively neutral universal service obligations.

2010/62. Thierry BRECHET, Julien THENIE, Thibaut ZEIMES and Stéphane ZUBER. The benefits of cooperation under uncertainty: the case of climate change.

2010/63. Marco DI SUMMA and Laurence A. WOLSEY. Mixing sets linked by bidirected paths.

2010/64. Kaz MIYAGIWA, Huasheng SONG and Hylke VANDENBUSSCHE. Innovation, antidumping and retaliation.

2010/65. Thierry BRECHET, Natali HRITONENKO and Yuri YATSENKO. Adaptation and mitigation in long-term climate policies.

2010/66. Marc FLEURBAEY, Marie-Louise LEROUX and Gregory PONTHIERE. Compensating the dead? Yes we can!

2010/67. Philippe CHEVALIER, Jean-Christophe VAN DEN SCHRIECK and Ying WEI. Measuring the variability in supply chains with the peakedness.

2010/68. Mathieu VAN VYVE. Fixed-charge transportation on a path: optimization, LP formulations and separation.

2010/69. Roland Iwan LUTTENS. Lower bounds rule!

2010/70. Fred SCHROYEN and Adekola OYENUGA. Optimal pricing and capacity choice for a public service under risk of interruption.

2010/71. Carlotta BALESTRA, Thierry BRECHET and Stéphane LAMBRECHT. Property rights with biological spillovers: when Hardin meets Meade.

2010/72. Olivier GERGAUD and Victor GINSBURGH. Success: talent, intelligence or beauty?

2010/73. Jean GABSZEWICZ, Victor GINSBURGH, Didier LAUSSEL and Shlomo WEBER. Foreign languages' acquisition: self learning and linguistic schools.

2010/74. Cédric CEULEMANS, Victor GINSBURGH and Patrick LEGROS. Rock and roll bands, (in)complete contracts and creativity.

2010/75. Nicolas GILLIS and François GLINEUR. Low-rank matrix approximation with weights or missing data is NP-hard.

2010/76. Ana MAULEON, Vincent VANNETELBOSCH and Cecilia VERGARI. Unions' relative concerns and strikes in wage bargaining.

2010/77. Ana MAULEON, Vincent VANNETELBOSCH and Cecilia VERGARI. Bargaining and delay in patent licensing.

Recent titles

CORE Discussion Papers - continued

2010/78. Jean J. GABSZEWICZ and Ornella TAROLA. Product innovation and market acquisition of firms.

2010/79. Michel LE BRETON, Juan D. MORENO-TERNERO, Alexei SAVVATEEV and Shlomo WEBER. Stability and fairness in models with a multiple membership.

2010/80. Juan D. MORENO-TERNERO. Voting over piece-wise linear tax methods.

2010/81. Jean HINDRIKS, Marijn VERSCHELDE, Glenn RAYP and Koen SCHOORS. School tracking, social segregation and educational opportunity: evidence from Belgium.

2010/82. Jean HINDRIKS, Marijn VERSCHELDE, Glenn RAYP and Koen SCHOORS. School autonomy and educational performance: within-country evidence.

2010/83. Dunia LOPEZ-PINTADO. Influence networks.

2010/84. Per AGRELL and Axel GAUTIER. A theory of soft capture.

2010/85. Per AGRELL and Roman KASPERZEC. Dynamic joint investments in supply chains under information asymmetry.

2010/86. Thierry BRECHET and Pierre M. PICARD. The economics of airport noise: how to manage markets for noise licenses.

2010/87. Eve RAMAEKERS. Fair allocation of indivisible goods among two agents.

2011/1. Yu. NESTEROV. Random gradient-free minimization of convex functions.

2011/2. Olivier DEVOLDER, François GLINEUR and Yu. NESTEROV. First-order methods of smooth convex optimization with inexact oracle.

Books

J. GABSZEWICZ (ed.) (2006), La différenciation des produits. Paris, La découverte.

L. BAUWENS, W. POHLMEIER and D. VEREDAS (eds.) (2008), *High frequency financial econometrics: recent developments*. Heidelberg, Physica-Verlag.

P. VAN HENTENRYCKE and L. WOLSEY (eds.) (2007), *Integration of AI and OR techniques in constraint programming for combinatorial optimization problems*. Berlin, Springer.

P-P. COMBES, Th. MAYER and J-F. THISSE (eds.) (2008), *Economic geography: the integration of regions and nations*. Princeton, Princeton University Press.

J. HINDRIKS (ed.) (2008), *Au-delà de Copernic: de la confusion au consensus* ? Brussels, Academic and Scientific Publishers.

J-M. HURIOT and J-F. THISSE (eds) (2009), *Economics of cities.* Cambridge, Cambridge University Press.

P. BELLEFLAMME and M. PEITZ (eds) (2010), *Industrial organization: markets and strategies*. Cambridge University Press.

M. JUNGER, Th. LIEBLING, D. NADDEF, G. NEMHAUSER, W. PULLEYBLANK, G. REINELT, G. RINALDI and L. WOLSEY (eds) (2010), *50 years of integer programming, 1958-2008: from the early years to the state-of-the-art*. Berlin Springer.

CORE Lecture Series

C. GOURIÉROUX and A. MONFORT (1995), Simulation Based Econometric Methods.

A. RUBINSTEIN (1996), Lectures on Modeling Bounded Rationality.

J. RENEGAR (1999), A Mathematical View of Interior-Point Methods in Convex Optimization.

B.D. BERNHEIM and M.D. WHINSTON (1999), Anticompetitive Exclusion and Foreclosure Through Vertical Agreements.

D. BIENSTOCK (2001), Potential function methods for approximately solving linear programming problems: theory and practice.

R. AMIR (2002), Supermodularity and complementarity in economics.

R. WEISMANTEL (2006), Lectures on mixed nonlinear programming.