2007/76

■

Gradient methods for minimizing composite objective function

Yu. Nesterov

CORE DISCUSSION PAPER
2007/76

Gradient methods for minimizing composite objective function

Yu. NESTEROV[1]

September 2007

Abstract

In this paper we analyze several new methods for solving optimization problems with the objective function formed as a sum of two convex terms: one is smooth and given by a black-box oracle, and another is general but simple and its structure is known. Despite to the bad properties of the sum, such problems, both in convex and nonconvex cases, can be solved with efficiency typical for the good part of the objective. For convex problems of the above structure, we consider primal and dual variants of the gradient method (converge as $O(1/k)$), and an accelerated multistep version with convergence rate $O(1/k^2)$, where k is the iteration counter. For all methods, we suggest some efficient "line search" procedures and show that the additional computational work necessary for estimating the unknown problem class parameters can only multiply the complexity of each iteration by a small constant factor. We present also the results of preliminary computational experiments, which confirm the superiority of the accelerated scheme.

Keywords: local optimization, convex optimization, nonsmooth optimization, complexity theory, black-box model, optimal methods, structural optimization, l_1-regularization.

---

[1] CORE and INMA, Université catholique de Louvain, Belgium. E-mail: yurii.nesterov@uclouvain.be. The author is also member of ECORE, the newly created association between CORE and ECARES.

The research results presented in this paper have been supported by a grant "Action de recherche concertée ARC 04/09-315" from the "Direction de la recherché scientifique – Communauté française de Belgique".

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 author.

1 Introduction

Motivation. In the last years, several advances in Convex Optimization were based on development of different *models* for optimization problems. Starting from the theory of self-concordant functions [12], it was becoming more and more clear that the proper use of the problem's *structure* can lead to very efficient optimization methods, which significantly overpass the limitations of the black-box Complexity Theory (see Section 4.1 in [8] for discussion). For the recent examples, we can mention the development of *smoothing technique* [9], or the special methods for minimizing convex objective function up to certain *relative accuracy* [10]. In both cases, the proposed optimization schemes strongly employ the particular structure of corresponding optimization problem.

In this paper, we develop new optimization methods for approximating a global minimum of *composite* convex objective function $\phi(x)$. Namely, we assume that

$$\phi(x) \;=\; f(x) + \Psi(x), \tag{1.1}$$

where $f(x)$ is a differentiable convex function defined by a black-box oracle, and $\Psi(x)$ is a general closed convex function. However, we assume that function $\Psi(x)$ is *simple*. This means that we are able to find a closed-form solution for minimizing the sum of Ψ with some simple auxiliary functions. Let us give several examples.

1. Constrained minimization. Let Q be a closed convex set. Define Ψ as an indicator function of the set Q:

$$\Psi(x) \;=\; \begin{cases} 0, & \text{if } x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Then, the unconstrained minimization of composite function (1.1) is equivalent to minimizing function f over the set Q. We will see that our assumption on simplicity of function Ψ reduces to ability of finding in a closed form a Euclidean projection of arbitrary point onto the set Q.

2. Barrier representation of feasible set. Assume that the objective function of convex constrained minimization problem

$$\text{find } f^* \;=\; \min_{x \in Q} f(x)$$

is given by a black-box oracle, but the feasible set Q is described by a ν-self-concordant barrier $F(x)$ [12]. Define $\Psi(x) = \frac{\epsilon}{\nu} F(x)$, $\phi(x) = f(x) + \Psi(x)$, and $x^* \;=\; \arg\min_{x \in Q} f(x)$. Then, for arbitrary $\hat{x} \in \operatorname{int} Q$, by general properties of self-concordant barriers we get

$$\begin{aligned} f(\hat{x}) \;&\leq\; f(x^*) + \langle \nabla \phi(\hat{x}), \hat{x} - x^* \rangle + \tfrac{\epsilon}{\nu} \langle \nabla F(\hat{x}), x^* - \hat{x} \rangle \\ &\leq\; f^* + \|\nabla \phi(\hat{x})\|^* \cdot \|\hat{x} - x^*\| + \epsilon. \end{aligned}$$

Thus, a point $\hat{x}$, with small norm of the gradient of function ϕ, approximates well the solution of the constrained minimization problem. Note that the objective function ϕ does not belong to any standard class of convex problems formed by functions with bounded derivatives of certain degree.

3. Sparse least squares. In many applications, it is necessary to minimize the following objective:

$$\phi(x) \;=\; \tfrac{1}{2}\|Ax-b\|_2^2 + \|x\|_1 \;\stackrel{\rm def}{=}\; f(x)+\Psi(x), \tag{1.2}$$

where A is a matrix of corresponding dimension and $\|\cdot\|_k$ denotes the standard l_k-norm. The presence of additive l_1-term very often increases the sparsity of the optimal solution (see [1, 16]). This feature was observed a long time ago (see, for example, [2, 5, 14, 15]). Recently, this technique became popular in signal processing and statistics [6, 17].[1)]

From the formal point of view, the objective $\phi(x)$ in (1.2) is a nonsmooth convex function. Hence, the standard black-box gradient schemes need $O(\frac{1}{\epsilon^2})$ iterations for generating its ϵ-solution. The structural methods based on the smoothing technique [9] need $O(\frac{1}{\epsilon})$ iterations. However, we will see that the same problem can be solved in $O(\frac{1}{\epsilon^{1/2}})$ iterations of a special gradient-type scheme.

Contents. In Section 2 we introduce the *composite gradient mapping*. Its objective function is formed as a sum of objective of the usual gradient mapping [7] and the general nonsmooth convex term Ψ. For the particular case (1.2), this construction was proposed in [18]. In this section, we present different properties of this object, which are important for complexity analysis of optimization methods. In Section 3 we study the behavior of the simplest gradient scheme based on the composite gradient mapping. We prove that in convex and nonconvex cases we have exactly the same complexity results as in the usual smooth situation ($\Psi \equiv 0$). For example, in the case of convex f with Lipschitz continuous gradient, the Gradient Method converges as $O(\frac{1}{k})$, where k is the iteration counter. It is important that our version of the Gradient Method has an adjustable stepsize strategy, which needs in average one additional computation of the function value per iteration.

In the next Section 4, we introduce a machinery of estimate sequences and apply it first for justifying the rate of convergence of the dual variant of the gradient method. After, we present an accelerated version, which converges as $O(\frac{1}{k^2})$. As compared with the previous variants of accelerated schemes (e.g. [8], [9]), our new scheme can efficiently adjust the initial estimate of the unknown Lipschitz constant. In Section 5 we give examples of applications of the accelerated scheme. We show how to minimize functions with known strong convexity parameter (Section 5.1), how to find a point with a small residual in the system of the first-order optimality conditions (Section 5.2), and how to approximate unknown parameter of strong convexity (Section 5.3). In the last Section 6 we present the results of preliminary testing of the proposed optimization methods.

Notation. In what follows E, denotes a finite-dimensional real vector space, and E^* the dual space, which is formed by all linear functions on E. The value of function $s \in E^*$ at $x \in E$ is denoted by $\langle s, x\rangle$. By fixing a positive definite self-adjoint operator $B : E \to E^*$, we can define the following Euclidean norms:

$$\begin{aligned} \|h\| &= \langle Bh, h\rangle^{1/2}, \quad h \in E, \\ \|s\|_* &= \langle s, B^{-1}s\rangle^{1/2}, \quad s \in E^*. \end{aligned} \tag{1.3}$$

[1)] An interested reader can find a good survey of the literature, existing minimization technique, and new methods in [3] and [4].

In particular case of coordinate vector space $E = R^n$, we have $E = E^*$. Then, usually B is taken as a unit matrix, and $\langle s, x \rangle$ denotes the standard coordinate-wise inner product.

Further, for function $f(x)$, $x \in E$, we denote by $\nabla f(x)$ its *gradient* at x:

$$f(x+h) \;=\; f(x) + \langle \nabla f(x), h \rangle + o(\|h\|), \quad h \in E.$$

Clearly $\nabla f(x) \in E^*$. For convex function Ψ we denote by $\partial \Psi(x)$ its subdifferential at x. Finally, the directional derivative of function ϕ is defined in the usual way:

$$D\phi(y)[u] \;=\; \lim_{\alpha \downarrow 0} \tfrac{1}{\alpha}[\phi(y + \alpha u) - \phi(y)].$$

2 Composite gradient mapping

In this paper, we consider a problem of approximating a local minimum of function

$$\phi(x) \;\stackrel{\rm def}{=}\; f(x) + \Psi(x) \tag{2.1}$$

over a convex set Q, where function f is differentiable, and function Ψ is closed and convex on Q. For characterizing a solution to our problem, define the cone of *feasible directions* and the corresponding dual cone, which is called *normal*:

$$\begin{aligned} \mathcal{F}(y) &= \{u = \tau \cdot (x - y),\ x \in Q,\ \tau \geq 0\} \subseteq E, \\ \mathcal{N}(y) &= \{s :\ \langle s, x - y \rangle \geq 0,\ x \in Q\} \subseteq E^*, \quad y \in Q. \end{aligned}$$

Then, the first-order necessary optimality conditions at the point of local minimum x^* can be written as follows:

$$\phi'_* \;\stackrel{\rm def}{=}\; \nabla f(x^*) + \xi^* \;\in\; \mathcal{N}(x^*), \tag{2.2}$$

where $\xi^* \in \partial \Psi(x^*)$. In other words,

$$\langle \phi'_*, u \rangle \;\geq\; 0 \quad \forall u \in \mathcal{F}(x^*). \tag{2.3}$$

Since Ψ is convex, the latter condition is equivalent to the following:

$$D\phi(x^*)[u] \;\geq\; 0 \quad \forall u \in \mathcal{F}(x^*). \tag{2.4}$$

Note that in the case of convex f, any of the conditions (2.2) - (2.4) is sufficient for point x^* to be a point of global minimum of function ϕ over Q.

The last variant of the first-order optimality conditions is convenient for defining an approximate solution to our problem.

Definition 1 *The point $\bar{x} \in Q$ satisfies the first-order optimality conditions of local minimum of function ϕ over the set Q with accuracy $\epsilon \geq 0$ if*

$$D\phi(\bar{x})[u] \;\geq\; -\epsilon \quad \forall u \in \mathcal{F}(\bar{x}),\ \|u\| = 1. \tag{2.5}$$

Note that in the case $\mathcal{F}(\bar{x}) = E$ with $0 \notin \nabla f(\bar{x}) + \partial\Psi(\bar{x})$, this condition reduces to the following inequality:

$$\begin{aligned} -\epsilon &\leq \min_{\|u\|=1} D\phi(\bar{x})[u] = \min_{\|u\|=1} \max_{\xi\in\partial\Psi(\bar{x})} \langle \nabla f(\bar{x}) + \xi, u\rangle \\ &= \min_{\|u\|\leq 1} \max_{\xi\in\partial\Psi(\bar{x})} \langle \nabla f(\bar{x}) + \xi, u\rangle = \max_{\xi\in\partial\Psi(\bar{x})} \min_{\|u\|\leq 1} \langle \nabla f(\bar{x}) + \xi, u\rangle \\ &= -\min_{\xi\in\partial\Psi(\bar{x})} \|\nabla f(\bar{x}) + \xi\|_*. \end{aligned}$$

For finding a point $\bar{x}$ satisfying condition (2.5), we are going to use the *composite gradient mapping*. Namely, at any $y \in Q$ define

$$\begin{aligned} m_L(y; x) &= f(y) + \langle \nabla f(y), x - y\rangle + \tfrac{L}{2}\|x - y\|^2 + \Psi(x), \\ T_L(y) &= \arg\min_{x\in Q} m_L(y; x), \end{aligned} \tag{2.6}$$

where L is a positive constant.[2] Then, we can define a constrained analogue of the gradient direction of smooth function, the vector

$$g_L(y) = L \cdot B(y - T_L(y)) \in E^*. \tag{2.7}$$

(In case of an ambiguity with objective function, we use notation $g_L(y)[\phi]$.) It is easy to see that for $Q \equiv E$ and $\Psi \equiv 0$ we get $g_L(y) = \nabla\phi(y) \equiv \nabla f(x)$ for any $L > 0$. Our assumption on simplicity of function Ψ means exactly the feasibility of operation (2.6).

Let us mention the main properties of the composite gradient mapping. Almost all of them follow from the first-order optimality condition for problem (2.6):

$$\langle \nabla f(y) + LB(T_L(y) - y) + \xi_L(y), x - T_L(y)\rangle \geq 0, \quad \forall x \in Q, \tag{2.8}$$

where $\xi_L(y) \in \partial\Psi(T_L(y))$. In what follows, we denote

$$\phi'(T_L(y)) = \nabla f(T_L(y)) + \xi_L(y) \in \partial\phi(T_L(y)). \tag{2.9}$$

We are going to show that the above subgradient inherits all important properties of the gradient of smooth convex function.

From now on, we assume that the first part of the objective function (2.1) has Lipschitz-continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_f\|x - y\|, \quad x, y \in Q, \tag{2.10}$$

From (2.10) and convexity of Q, one can easily derive the following useful inequality (see, for example, [13]):

$$|f(x) - f(y) - \langle \nabla f(y), x - y\rangle| \leq \tfrac{L_f}{2}\|x - y\|^2, \quad x, y \in Q. \tag{2.11}$$

First of all, let us estimate a local variation of function ϕ. Denote

$$S_L(y) = \tfrac{\|\nabla f(T_L(y)) - \nabla f(y)\|_*}{\|T_L(y) - y\|} \leq L_f.$$

[2] Recall that in the usual gradient mapping [7] we have $\Psi(\cdot) \equiv 0$. Our modification is inspired by [18].

Theorem 1 *At any $y \in Q$,*

$$\begin{aligned} \phi(y) - \phi(T_L(y)) &\geq \tfrac{2L-L_f}{2L^2}\|g_L(y)\|_*^2, & (2.12)\\ \langle \phi'(T_L(y)), y - T_L(y)\rangle &\geq \tfrac{L-L_f}{L^2}\|g_L(y)\|_*^2. & (2.13) \end{aligned}$$

Moreover, for any $x \in Q$, we have

$$\begin{aligned} \langle \phi'(T_L(y)), x - T_L(y)\rangle &\geq -\left(1 + \tfrac{1}{L}S_L(y)\right)\cdot\|g_L(y)\|_*\cdot\|T_L(y) - x\| \\ &\geq -\left(1 + \tfrac{L_f}{L}\right)\cdot\|g_L(y)\|_*\cdot\|T_L(y) - x\|. \end{aligned} \quad (2.14)$$

Proof:
For the sake of notation, denote $T = T_L(y)$ and $\xi = \xi_L(y)$. Then

$$\begin{aligned} \phi(T) &\overset{(2.10)}{\leq} f(y) + \langle \nabla f(y), T - y\rangle + \tfrac{L_f}{2}\|T - y\|^2 + \Psi(T) \\ &\overset{(2.8),\,x=y}{\leq} f(y) + \langle LB(T - y) + \xi, y - T\rangle + \tfrac{L_f}{2}\|T - y\|^2 + \Psi(T) \\ &= f(y) + \tfrac{L_f - 2L}{2}\|T - y\|^2 + \Psi(T) + \langle \xi, y - T\rangle \\ &\leq \phi(y) - \tfrac{2L - L_f}{2}\|T - y\|^2. \end{aligned}$$

Taking into account the definition (2.7), we get (2.12). Further,

$$\begin{aligned} \langle \nabla f(T) + \xi, y - T\rangle &= \langle \nabla f(y) + \xi, y - T\rangle - \langle \nabla f(T) - \nabla f(y), T - y\rangle \\ &\overset{(2.8),\,x=y}{\geq} \langle LB(y - T), y - T\rangle - \langle \nabla f(T) - \nabla f(y), T - y\rangle \\ &\overset{(2.10)}{\geq} (L - L_f)\|T - y\|^2 \overset{(2.7)}{=} \tfrac{L-L_f}{L^2}\|g_L(y)\|_*^2. \end{aligned}$$

Thus, we get (2.13). Finally,

$$\begin{aligned} \langle \nabla f(T) + \xi, T - x\rangle &\overset{(2.8)}{\leq} \langle \nabla f(T), T - x\rangle + \langle \nabla f(y) + LB(T - y), x - T\rangle \\ &= \langle \nabla f(T) - \nabla f(y), T - x\rangle - \langle g_L(y), x - T\rangle \\ &\overset{(2.7)}{\leq} \left(1 + \tfrac{1}{L}S_L(y)\right)\cdot\|g_L(y)\|_*\cdot\|T - x\|, \end{aligned}$$

and (2.14) follows. □

Corollary 1 *For any $y \in Q$, and any $u \in \mathcal{F}(T_L(y))$, $\|u\| = 1$, we have*

$$D\phi(T_L(y))[u] \geq -\left(1 + \tfrac{L_f}{L}\right)\cdot\|g_L(y)\|_*. \quad (2.15)$$

In this respect, it is interesting to investigate the dependence of $\|g_L(y)\|_*$ in L.

Lemma 1 *The norm of the gradient direction $\|g_L(y)\|_*$ is increasing in L, and the norm of the step $\|T_L(y) - y\|$ is decreasing in L.*

Proof:
Indeed, consider the function

$$\omega(\tau) \;=\; \min_{x \in Q} \left[f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{1}{2\tau} \|x - y\|^2 + \Psi(x) \right].$$

The objective function of this minimization problem is jointly convex in x and τ. Therefore, $\omega(\tau)$ is convex in τ. Since the minimum of this problem is attained at a single point, $\omega(\tau)$ is differentiable and

$$\omega'(\tau) \;=\; -\tfrac{1}{2} \|\tfrac{1}{\tau}[T_{1/\tau}(y) - y]\|^2 \;=\; -\tfrac{1}{2} \|g_{1/\tau}(y)\|_*^2.$$

Since $\omega(\cdot)$ is convex, $\omega'(\tau)$ is an increasing function of τ. Hence, $\|g_{1/\tau}(y)\|_*$ is a decreasing function of τ.

For the second statement follows from concavity of function

$$\hat{\omega}(L) \;=\; \min_{x \in Q} \left[f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{L}{2} \|x - y\|^2 + \Psi(x) \right].$$

□

Now let us look at the output of the composite gradient mapping from a global perspective.

Theorem 2 *For any $y \in Q$ we have*

$$m_L(y; T_L(y)) \;\leq\; \phi(y) - \tfrac{1}{2L} \|g_L(y)\|_*^2, \tag{2.16}$$

$$m_L(y; T_L(y)) \;\leq\; \min_{x \in Q} \left[\phi(x) + \tfrac{L + L_f}{2} \|x - y\|^2 \right]. \tag{2.17}$$

If function f is convex, then

$$m_L(y; T_L(y)) \;\leq\; \min_{x \in Q} \left[\phi(x) + \tfrac{L}{2} \|x - y\|^2 \right]. \tag{2.18}$$

Proof:
Note that function $m_L(y; x)$ is strongly convex in x with convexity parameter L. Hence,

$$\phi(y) - m_L(y; T_L(y)) \;=\; m_L(y; y) - m_L(y; T_L(y)) \;\geq\; \tfrac{L}{2} \|y - T_L(y)\|^2 \;=\; \tfrac{1}{2L} \|g_L(y)\|_*^2.$$

Further, if f is convex, then

$$\begin{aligned} m_L(y; T_L(y)) \;&=\; \min_{x \in Q} \left[f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{L}{2} \|x - y\|^2 + \Psi(x) \right] \\ &\leq\; \min_{x \in Q} \left[f(x) + \Psi(x) + \tfrac{L}{2} \|x - y\|^2 \right] \\ &=\; \min_{x \in Q} \left[\phi(x) + \tfrac{L}{2} \|x - y\|^2 \right]. \end{aligned}$$

For nonconvex f, we can plug into the same reasoning the following consequence of (2.11):

$$f(y) + \langle \nabla f(y), x - y \rangle \;\; \leq \;\; f(x) + \tfrac{L_f}{2} \|x - y\|^2.$$

□

Remark 1 *In view of (2.10), for $L \geq L_f$ we have*

$$\phi(T_L(y)) \;\; \leq \;\; m_L(y; T_L(y)). \tag{2.19}$$

Hence, in this case inequality (2.18) guarantees

$$\phi(T_L(y)) \;\; \leq \;\; \min_{x \in Q} \left[\phi(x) + \tfrac{L}{2} \|x - y\|^2 \right]. \tag{2.20}$$

Finally, let us prove a useful inequality for strongly convex ϕ.

Lemma 2 *Let function ϕ be strongly convex with convexity parameter $\mu_\phi > 0$. Then for any $y \in Q$ we have*

$$\|T_L(y) - x^*\| \;\; \leq \;\; \tfrac{1}{\mu_\phi} \cdot \left(1 + \tfrac{1}{L} S_L(y)\right) \cdot \|g_L(y)\|_* \;\; \leq \;\; \tfrac{1}{\mu_\phi} \cdot \left(1 + \tfrac{L_f}{L}\right) \cdot \|g_L(y)\|_*, \tag{2.21}$$

where x^* is a unique minimum of ϕ on Q.

Proof:
Indeed, in view of inequality (2.14), we have:

$$\begin{aligned} \left(1 + \tfrac{L_f}{L}\right) \cdot \|g_L(y)\|_* \cdot \|T_L(y) - x^*\| \;\; &\geq \;\; \left(1 + \tfrac{1}{L} S_L(y)\right) \cdot \|g_L(y)\|_* \cdot \|T_L(y) - x^*\| \\ &\geq \;\; \langle \phi'(T_L(y)), T_L(y) - x^* \rangle \;\; \geq \;\; \mu_\phi \|T_L(y) - x^*\|^2, \end{aligned}$$

and (2.21) follows. □

Now we are ready to analyze different optimization schemes based on the composite gradient mapping. In the next section, we describe the simplest one.

3 Gradient method

Define first the gradient iteration with the simplest backtracking strategy for the "line search" parameter (we call its termination condition the *full relaxation*).

Gradient Iteration $\mathcal{G}(x, M)$

SET:	$L := M.$
REPEAT:	$T := T_L(x),$
	if $\phi(T) > m_L(x; T)$ **then** $L := L \cdot \gamma_u,$
UNTIL:	$\phi(T) \leq m_L(x; T).$
OUTPUT:	$\mathcal{G}(x, M).T \;=\; T, \quad \mathcal{G}(x, M).L \;=\; L,$
	$\mathcal{G}(x, M).S \;=\; S_L(x).$

(3.1)

If there exists an ambiguity in the objective function, we use notation $\mathcal{G}_\phi(x, M)$.

For running the gradient scheme, we need to choose an initial optimistic estimate L_0 for the Lipschitz constant L_f:

$$0 \;<\; L_0 \;\leq\; L_f, \tag{3.2}$$

and two adjustment parameters $\gamma_u > 1$ and $\gamma_d \geq 1$. Let $y_0 \in Q$ be our starting point. For $k \geq 0$, consider the following iterative process.

Gradient Method $\mathcal{GM}(y_0, L_0)$

$$\begin{aligned} y_{k+1} &= \mathcal{G}(y_k, L_k).T, \\ M_k &= \mathcal{G}(y_k, L_k).L, \\ L_{k+1} &= \max\{L_0, M_k/\gamma_d\}. \end{aligned} \tag{3.3}$$

Thus, $y_{k+1} = T_{M_k}(y_k)$. Since function f satisfies inequality (2.11), in the loop (3.1), the value L can keep increasing only if $L \leq L_f$. Taking into account condition (3.2), we obtain the following bounds:

$$L_0 \;\leq L_k \;\leq\; M_k \;\leq\; \gamma_u L_f. \tag{3.4}$$

Moreover, if $\gamma_d \geq \gamma_u$, then

$$L_k \;\leq\; L_f, \quad k \geq 0. \tag{3.5}$$

Note that in (3.1) there is no explicit bound on the number of repetition of the loop. However, it is easy to see that the total amount of calls of oracle N_k after k iterations of (3.3) cannot be too big.

Lemma 3 *In the method (3.3), for any $k \geq 0$ we have*

$$N_k \;\leq\; \left[1 + \tfrac{\ln \gamma_d}{\ln \gamma_u}\right] \cdot (k+1) + \tfrac{1}{\ln \gamma_u} \cdot \left(\ln \tfrac{\gamma_u L_f}{\gamma_d L_0}\right)_+. \tag{3.6}$$

Proof:
Denote by $n_i \geq 1$ the number of calls of the oracle at iteration $i \geq 0$. Then

$$L_{i+1} \;\geq\; \tfrac{1}{\gamma_d} \cdot L_i \cdot \gamma_u^{n_i - 1}.$$

Thus,

$$n_i \;\leq\; 1 + \tfrac{\ln \gamma_d}{\ln \gamma_u} + \tfrac{1}{\ln \gamma_u} \cdot \ln \tfrac{L_{i+1}}{L_i}.$$

Hence, we can estimate

$$N_k \;\leq\; \sum_{i=0}^{k} n_i \;=\; \left[1 + \tfrac{\ln \gamma_d}{\ln \gamma_u}\right] \cdot (k+1) + \tfrac{1}{\ln \gamma_u} \cdot \ln \tfrac{L_{k+1}}{L_0}.$$

In remains to note that $L_{k+1} \overset{(3.4)}{\leq} \max\left\{L_0, \tfrac{\gamma_u}{\gamma_d} L_f\right\}$. □

A reasonable choice of the adjustment parameters is as follows:

$$\gamma_u \;=\; \gamma_d \;=\; 2 \quad \overset{(3.6)}{\Rightarrow} \quad N_k \;\leq\; 2(k+1) + \log_2 \tfrac{L_f}{L_0}, \quad L_k \;\overset{(3.5)}{\leq}\; L_f. \tag{3.7}$$

Thus, the performance of the Gradient Method (3.3) is well described by the estimates for the iteration counter, Therefore, in the rest part of this section we will focus on estimating the rate of convergence of this method in different situations.

Let us start from the general nonconvex case. Denote

$$\delta_k \;=\; \min_{0 \leq i \leq k} \tfrac{1}{2M_i} \|g_{M_i}(y_i)\|_*^2,$$

$$i_k \;=\; 1 + \arg\min_{0 \leq i \leq k} \tfrac{1}{2M_i} \|g_{M_i}(y_i)\|_*^2.$$

Theorem 3 *Let function ϕ be bounded below on Q by some constant ϕ_*. Then*

$$\delta_k \;\leq\; \tfrac{\phi(y_0) - \phi^*}{k+1}. \tag{3.8}$$

Moreover, for any $u \in \mathcal{F}(y_{i_k})$ with $\|u\| = 1$ we have

$$D\phi(y_{i_k})[u] \;\geq\; -\tfrac{(1+\gamma_u) L_f}{L_0^{1/2}} \cdot \sqrt{\tfrac{2(\phi(y_0) - \phi_*)}{k+1}}. \tag{3.9}$$

Proof:
Indeed, in view of the termination criterion in (3.1), we have

$$\phi(y_i) - \phi(y_{i+1}) \;\geq\; \phi(y_i) - m_{M_i}(y_i; T_{M_i}(y_i)) \overset{(2.16)}{\geq} \tfrac{1}{2M_i}\|g_{M_i}(y_i)\|_*^2.$$

Summing up these inequalities for $i = 0, \ldots, k$, we obtain (3.8).

Denote $j_k = i_k - 1$. Since $y_{i_k} = T_{M_{j_k}}(y_{j_k})$, for any $u \in \mathcal{F}(y_{i_k})$ with $\|u\| = 1$ we have

$$D\phi(y_{i_k})[u] \overset{(2.15)}{\geq} -\left(1 + \tfrac{L_f}{M_{j_k}}\right) \cdot \|g_{M_{j_k}}(y_{j_k})\|_* \;=\; -\left(1 + \tfrac{L_f}{M_{j_k}}\right) \cdot \sqrt{2M_{j_k}\delta_k}$$

$$\overset{(3.8)}{\geq} -\tfrac{M_{j_k}+L_f}{M_{j_k}^{1/2}} \cdot \sqrt{\tfrac{2(\phi(y_0)-\phi_*)}{k+1}} \overset{(3.4)}{\geq} -\tfrac{(1+\gamma_u)L_f}{L_0^{1/2}} \cdot \sqrt{\tfrac{2(\phi(y_0)-\phi_*)}{k+1}}.$$

□

Let us describe now the behavior of the Gradient Method (3.3) in convex case.

Theorem 4 *Let function f be convex on Q. Assume that it attains a minimum on Q at point x^* and that the level sets of ϕ are bounded:*

$$\|y - x^*\| \;\leq\; R \quad \forall y \in Q : \; \phi(y) \leq \phi(y_0). \tag{3.10}$$

If $\phi(y_0) - \phi(x^*) \geq \gamma_u L_f R^2$, then $\phi(y_1) - \phi(x^*) \leq \frac{\gamma_u L_f R^2}{2}$. Otherwise, for any $k \geq 0$ we have

$$\phi(y_k) - \phi(x^*) \;\leq\; \tfrac{2\gamma_u L_f R^2}{k+2}. \tag{3.11}$$

Moreover, for any $u \in \mathcal{F}(y_{i_k})$ with $\|u\| = 1$ we have

$$D\phi(y_{i_k})[u] \;\geq\; -\tfrac{4(1+\gamma_u)L_f R}{k+3} \cdot \sqrt{\tfrac{L_f}{L_0}}. \tag{3.12}$$

Proof:
Since $\phi(y_{k+1}) \leq \phi(y_k)$ for all $k \geq 0$, we have the bound $\|y_k - x^*\| \leq R$ valid for all generated points. Consider

$$y_k(\alpha) \;=\; \alpha x^* + (1-\alpha) y_k \;\in\; Q \quad \alpha \in [0, 1].$$

Then,

$$\phi(y_{k+1}) \;\leq\; m_{M_k}(y_k; T_{M_k}(y_k)) \overset{(2.18)}{\leq} \min_{y \in Q}\left[\phi(y) + \tfrac{M_k}{2}\|y - y_k\|^2\right]$$

$$(y = y_k(\alpha)) \;\leq\; \min_{0 \leq \alpha \leq 1}\left[\phi(\alpha x^* + (1-\alpha)y_k) + \tfrac{M_k \alpha^2}{2}\|y_k - x^*\|^2\right]$$

$$\overset{(3.4)}{\leq} \min_{0 \leq \alpha \leq 1}\left[\phi(y_k) - \alpha(\phi(y_k) - \phi(x^*)) + \tfrac{\gamma_u L_f R^2}{2} \cdot \alpha^2\right].$$

If $\phi(y_0) - \phi(x^*) \geq \gamma_u L_f R^2$, then the optimal solution of the latter optimization problem is $\alpha = 1$ and we get

$$\phi(y_1) - \phi(x^*) \;\leq\; \tfrac{\gamma_u L_f R^2}{2}.$$

Otherwise, the optimal solution is

$$\alpha \;=\; \tfrac{\phi(y_k)-\phi(x^*)}{\gamma_u L_f R^2} \;\leq\; \tfrac{\phi(y_0)-\phi(x^*)}{\gamma_u L_f R^2} \leq 1,$$

and we obtain

$$\phi(y_{k+1}) \;\leq\; \phi(y_k) - \tfrac{[\phi(y_k)-\phi(y^*)]^2}{2\gamma_u L_f R^2}. \tag{3.13}$$

From this inequality, denoting $\lambda_k = \frac{1}{\phi(y_k)-\phi(x^*)}$, we get

$$\lambda_{k+1} \;\geq\; \lambda_k + \tfrac{\lambda_{k+1}}{2\lambda_k \gamma_u L_f R^2} \;\geq\; \lambda_k + \tfrac{1}{2\gamma_u L_f R^2}.$$

Hence, for $k \geq 0$ we have

$$\lambda_k \;\geq\; \tfrac{1}{\phi(y_0)-\phi(x^*)} + \tfrac{k}{2\gamma_u L_f R^2} \;\geq\; \tfrac{k+2}{2\gamma_u L_f R^2}.$$

Further, let us fix an integer m, $0 < m < k$. Since

$$\phi(y_i) - \phi(y_{i+1}) \;\geq\; \tfrac{1}{2M_i}\|g_{M_i}(y_i)\|_*^2, \quad i = 0, \ldots, k,$$

we have

$$(k-m+1)\delta_k \;\leq\; \sum_{i=m}^{k} \tfrac{1}{2M_i}\|g_{M_i}(y_i)\|_*^2 \;\leq\; \phi(y_m) - \phi(y_{k+1})$$

$$\leq\; \phi(y_m) - \phi(x^*) \overset{(3.11)}{\leq} \tfrac{2\gamma_u L_f R^2}{m+2}.$$

Denote $j_k = i_k - 1$. Then, for any $u \in \mathcal{F}(y_{i_k})$ with $\|u\| = 1$, we have

$$D\phi(y_{i_k})[u] \overset{(2.15)}{\geq} -\left(1 + \tfrac{L_f}{M_{j_k}}\right) \cdot \|g_{M_{j_k}}(y_{j_k})\|_* \;=\; -\left(1 + \tfrac{L_f}{M_{j_k}}\right) \cdot \sqrt{2M_{j_k}\delta_k}$$

$$\overset{(3.11)}{\geq} -2\tfrac{M_{j_k}+L_f}{M_{j_k}^{1/2}} \cdot \sqrt{\tfrac{\gamma_u L_f R^2}{(m+2)(k+1-m)}}$$

$$\overset{(3.4)}{\geq} -2(1+\gamma_u) L_f R \cdot \sqrt{\tfrac{L_f}{L_0(m+2)(k+1-m)}}.$$

Choosing $m = \lfloor \frac{k}{2} \rfloor$, we get $(m+2)(k+1-m) \geq \left(\frac{k+3}{2}\right)^2$. □

Theorem 5 *Let function ϕ be strongly convex on Q with convexity parameter μ_ϕ. If $\frac{\mu_\phi}{L_f} \geq 2\gamma_u$, then for any $k \geq 0$ we have*

$$\phi(y_k) - \phi(x^*) \;\leq\; \left(\tfrac{\gamma_u L_f}{\mu_\phi}\right)^k (\phi(y_0) - \phi(y^*)) \;\leq\; \tfrac{1}{2^k}(\phi(y_0) - \phi(y^*)). \tag{3.14}$$

Otherwise,

$$\phi(y_k) - \phi(x^*) \;\leq \left(1 - \tfrac{\mu_\phi}{4\gamma_u L_f}\right)^k \cdot (\phi(y_0) - \phi(y^*)). \tag{3.15}$$

Proof:
Since ϕ is strongly convex, for any $k \geq 0$ we have

$$\phi(y_k) - \phi(x^*) \quad \geq \quad \tfrac{\mu_\phi}{2}\|y_k - x^*\|^2. \tag{3.16}$$

Denote $y_k(\alpha) = \alpha x^* + (1-\alpha)y_k \in Q$, $\alpha \in [0,1]$. Then,

$$\begin{aligned}
\phi(y_{k+1}) &\overset{(2.18)}{\leq} \min_{0\leq\alpha\leq 1}\left[\phi(\alpha x^* + (1-\alpha)y_k) + \tfrac{M_k\alpha^2}{2}\|y_k - x^*\|^2\right] \\
&\overset{(3.4)}{\leq} \min_{0\leq\alpha\leq 1}\left[\phi(y_k) - \alpha(\phi(y_k) - \phi(x^*)) + \tfrac{\gamma_u L_f}{2}\cdot\alpha^2\|y_k - x^*\|^2\right] \\
&\overset{(3.16)}{\leq} \min_{0\leq\alpha\leq 1}\left[\phi(y_k) - \alpha\left(1 - \alpha\cdot\tfrac{\gamma_u L_f}{\mu_\phi}\right)(\phi(y_k) - \phi(x^*))\right].
\end{aligned}$$

The minimum of the last expression is achieved for $\alpha^* = \min\left\{1, \frac{\mu_\phi}{2\gamma_u L_f}\right\}$. Hence, if $\frac{\mu_\phi}{2\gamma_u L_f} \geq 1$, then $\alpha^* = 1$ and we get

$$\phi(y_{k+1}) - \phi(x^*) \quad \leq \quad \tfrac{\gamma_u L_f}{\mu_\phi}(\phi(y_k) - \phi(y^*)) \;\leq\; \tfrac{1}{2}(\phi(y_k) - \phi(y^*)).$$

If $\frac{\mu_\phi}{2\gamma_u L_f} \leq 1$, then $\alpha^* = \frac{\mu_\phi}{2\gamma_u L_f}$ and

$$\phi(y_{k+1}) - \phi(x^*) \quad \leq \left(1 - \tfrac{\mu_\phi}{4\gamma_u L_f}\right)\cdot(\phi(y_k) - \phi(y^*)).$$

□

Remark 2 *1) In Theorem 5, the "condition number" $\frac{L_f}{\mu_\phi}$ can be smaller than one.*
2) For strongly convex ϕ, the bounds on the directional derivatives can be obtained by combining the inequalities (3.14), (3.15) with the estimate

$$\phi(y_k) - \phi(x^*) \overset{(2.12):L=L_f}{\geq} \tfrac{1}{2L_f}\|g_{L_f}(y_k)\|_*^2$$

and inequality (2.15). Thus, inequality (3.14) results in the bound

$$D\phi(y_{k+1})[u] \quad \geq \quad -2\left(\tfrac{\gamma_u L_f}{\mu_\phi}\right)^{k/2}\cdot\sqrt{2L_f(\phi(y_0) - \phi_*)}, \tag{3.17}$$

and inequality (3.15) leads to the bound

$$D\phi(y_{k+1})[u] \quad \geq \quad -2\left(1 - \tfrac{\mu_\phi}{4\gamma_u L_f}\right)^{k/2}\cdot\sqrt{2L_f(\phi(y_0) - \phi_*)}, \tag{3.18}$$

which are valid for all $u \in \mathcal{F}(y_{k+1})$ with $\|u\| = 1$.

4 Accelerated scheme

In the previous section, we have seen that, for convex f, the gradient method (3.3) converges as $O(\frac{1}{k})$. However, it is well known that on the convex problems the usual gradient scheme can be accelerated (e.g. Chapter 2 in [8]). Let us show that the same acceleration can be achieved for composite objective function.

Consider the problem

$$\min_{x \in E} \quad [\ \phi(x) = f(x) + \Psi(x)\ ], \tag{4.1}$$

where function f is convex and satisfies (2.10), and function Ψ is closed and strongly convex on E with convexity parameter $\mu_\Psi \geq 0$. We assume this parameter to be known. The case $\mu_\Psi = 0$ corresponds to convex Ψ. Denote by x^* the optimal solution to (4.1).

In problem (4.1), we allow $\mathrm{dom}\, \Psi \neq E$. Therefore, the formulation (4.1) covers also the constrained problems instances. Note that for (4.1), the first-order optimality conditions (2.8) defining the composite gradient mapping can be written in a simpler form:

$$\begin{aligned} T_L(y) &\in \mathrm{dom}\, \Psi, \\ \nabla f(y) + \xi_L(y) &= LB(y - T_L(y)) \ \equiv \ g_L(y), \end{aligned} \tag{4.2}$$

where $\xi_L(y) \in \partial \Psi(T_L(y))$.

For justifying the rate of convergence of different schemes as applied to (4.1), we will use the machinery of estimate functions in its newer variant [11]. Taking into account the special form of the objective in (4.1), we update recursively the following sequences.

- A minimizing sequence $\{x_k\}_{k=0}^{\infty}$.
- A sequence of increasing scaling coefficients $\{A_k\}_{k=0}^{\infty}$:

$$A_0 = 0, \quad A_k \stackrel{\mathrm{def}}{=} A_{k-1} + a_k, \quad k \geq 1.$$

- Sequence of estimate functions

$$\psi_k(x) \ = \ l_k(x) + A_k \Psi(x) + \tfrac{1}{2}\|x - x_0\|^2 \quad k \geq 0, \tag{4.3}$$

  where $x_0 \in \mathrm{dom}\, \Psi$ is our starting point, and $l_k(x)$ are linear functions in $x \in E$.

However, as compared with [11], we will add a possibility to update the estimates for Lipschitz constant L_f, using the initial guess L_0 satisfying (3.2), and two adjustment parameters $\gamma_u > 1$ and $\gamma_d \geq 1$.

For the above objects, we maintain recursively the following relations:

$$\left.\begin{aligned} \mathcal{R}_k^1: \quad A_k \phi(x_k) &\leq \psi_k^* \ \equiv \ \min_x \psi_k(x), \\ \mathcal{R}_k^2: \quad \psi_k(x) &\leq A_k \phi(x) + \tfrac{1}{2}\|x - x_0\|^2, \ \forall x \in E. \end{aligned}\right\}, \quad k \geq 0. \tag{4.4}$$

These relations clearly justify the following rate of convergence of the minimizing sequence:

$$\phi(x_k) - \phi(x^*) \ \leq \ \tfrac{\|x^* - x_0\|^2}{2A_k}, \quad k \geq 1. \tag{4.5}$$

Denote $v_k = \arg\min\limits_{x \in E} \psi_k(x)$. Since $\mu_{\psi_k} \geq 1$, for any $x \in E$ we have

$$A_k \phi(x_k) + \tfrac{1}{2}\|x - v_k\|^2 \overset{\mathcal{R}_k^1}{\leq} A_k \psi_k^* + \tfrac{1}{2}\|x - v_k\|^2 \leq \psi_k(x) \overset{\mathcal{R}_k^2}{\leq} A_k \phi(x) + \tfrac{1}{2}\|x - x_0\|^2.$$

Hence, taking $x = x^*$, we get two useful consequences of (4.4):

$$\|x^* - v_k\| \leq \|x^* - x_0\|, \qquad \|v_k - x_0\| \leq 2\|x^* - x_0\|, \quad k \geq 1. \tag{4.6}$$

Note that the relations (4.4) can be used for justifying the rate of convergence of a dual variant of the gradient method (3.3). Indeed, for $v_0 \in \operatorname{dom} \Psi$ define $\psi_0(x) = \frac{1}{2}\|x - v_0\|^2$, and choose L_0 satisfying condition (3.2).

Dual Gradient Method $\mathcal{DG}(v_0, L_0)$, $k \geq 0$.

$$\begin{aligned} y_k &= \mathcal{G}(v_k, L_k).T, \quad M_k = \mathcal{G}(v_k, L_k).L, \\ L_{k+1} &= \max\{L_0, M_k/\gamma_d\}, \quad a_{k+1} = \tfrac{1}{M_k}, \\ \psi_{k+1}(x) &= \psi_k(x) + \tfrac{1}{M_k}[f(v_k) + \langle \nabla f(v_k), x - v_k \rangle + \Psi(x)]. \end{aligned} \tag{4.7}$$

Since Ψ is simple, the points v_k are easily computable.

Note that the relations $\mathcal{R}_0^1$ and $\mathcal{R}_k^2$, $k \geq 0$, are trivial. Relations $\mathcal{R}_k^1$ can be justified by induction. Define $x_0 = y_0$, $\phi_k = \min\limits_{0 \leq i \leq k-1} \phi(y_i)$, and $x_k : \phi(x_k) = \phi_k$ for $k \geq 1$. Then

$$\begin{aligned} \psi_{k+1}^* &= \min_x \Big\{ \psi_k(x) + \tfrac{1}{M_k}[f(v_k) + \langle \nabla f(v_k), x - v_k \rangle + \Psi(x)] \Big\} \\ &\overset{\mathcal{R}_k^1}{\geq} A_k \phi_k + \min_x \Big\{ \tfrac{1}{2}\|x - v_k\|^2 + \tfrac{1}{M_k}[f(v_k) + \langle \nabla f(v_k), x - v_k \rangle + \Psi(x)] \Big\} \\ &\overset{(2.6)}{=} A_k \phi_k + a_{k+1} m_{M_k}(v_k; y_k) \\ &\overset{(3.1)}{\geq} A_k \phi_k + a_{k+1} \phi(y_k) \geq A_{k+1} \phi_{k+1}. \end{aligned}$$

Thus, relations $\mathcal{R}_k^1$ are valid for all $k \geq 0$. Since the values M_k satisfy bounds (3.4), for method (4.7) we obtain the following rate of convergence:

$$\phi(x_k) - \phi(x^*) \leq \tfrac{\gamma_u L_f}{2k}\|x^* - v_0\|^2, \quad k \geq 1. \tag{4.8}$$

Note that the constant in the right-hand side of this inequality is four times smaller than the constant in (3.11). However, each iteration in the dual method is two times more expensive as compared to the primal version (3.3).

However, the method (4.7) does not implement the best way of using the machinery of estimate functions. Let us look at the accelerated version of (4.7). As parameters, it

has the starting point $x_0 \in \mathrm{dom}\, \Psi$, the lower estimate $L_0 > 0$ for the Lipschitz constant L_f, and a lower estimate $\mu \in [0, \mu_\Psi]$ for the convexity parameter of function Ψ.

Accelerated method $\mathcal{A}(x_0, L_0, \mu)$

Initial settings: $\psi_0(x) = \frac{1}{2}\|x - x_0\|^2,\ A_0 = 0.$

Iteration $k \geq 0$

SET: $L := L_k$.

REPEAT: Find a from quadratic equation $\frac{a^2}{A_k + a} = 2\frac{1+\mu A_k}{L}$. $(*)$

Set $y = \frac{A_k x_k + a v_k}{A_k + a}$, and compute $T_L(y)$.

if $\langle \phi'(T_L(y)), y - T_L(y)\rangle < \frac{1}{L}\|\phi'(T_L(y))\|_*^2$**, then** $L := L \cdot \gamma_u$.

UNTIL: $\langle \phi'(T_L(y)), y - T_L(y)\rangle \geq \frac{1}{L}\|\phi'(T_L(y))\|_*^2$. $(**)$

DEFINE: $y_k := y,\ M_k := L,\ a_{k+1} := a,$

$L_{k+1} := M_k/\gamma_d,\ x_{k+1} := T_{M_k}(y_k),$

$\psi_{k+1}(x) := \psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)].$

(4.9)

As compared with Gradient Iteration (3.1), we use a *damped relaxation condition* $(**)$ as a stopping criterion of the internal cycle of (4.9).

Lemma 4 *Condition (**) in (4.9) is satisfied for any* $L \geq L_f$.

Proof:
Denote $T = T_L(y)$. Multiplying the representation

$$\phi'(T) \;=\; \nabla f(T) + \xi_L(y) \;\overset{(4.2)}{=}\; LB(y - T) + \nabla f(T) - \nabla f(y) \tag{4.10}$$

by vector $y - T$, we obtain

$$\begin{aligned}
\langle \phi'(T), y - T\rangle \;&=\; L\|y - T\|^2 - \langle \nabla f(y) - \nabla f(T), y - T\rangle \\
&\overset{(4.10)}{=}\; \tfrac{1}{L}\left[\|\phi'(T)\|^2 + 2L\langle \nabla f(y) - \nabla f(T), y - T\rangle - \|\nabla f(y) - \nabla f(T)\|_*^2\right] \\
&\quad -\langle \nabla f(y) - \nabla f(T), y - T\rangle \\
&=\; \tfrac{1}{L}\|\phi'(T)\|^2 + \langle \nabla f(y) - \nabla f(T), y - T\rangle - \tfrac{1}{L}\|\nabla f(y) - \nabla f(T)\|_*^2.
\end{aligned}$$

Hence, for $L \geq L_f$ condition (**) is satisfied. $\square$

Thus, we can always guarantee

$$L_k \;\leq\; M_k \;\leq\; \gamma_u L_f. \tag{4.11}$$

If $\gamma_d \geq \gamma_u$, then the upper bound (3.5) remains valid.

Let us establish a relation between the total number of calls of oracle N_k after k iterations, and the value of the iteration counter.

Lemma 5 *In the method (4.9), for any $k \geq 0$ we have*

$$N_k \;\leq\; 2\left[1 + \tfrac{\ln \gamma_d}{\ln \gamma_u}\right] \cdot (k+1) + \tfrac{1}{\ln \gamma_u} \cdot \ln \tfrac{2\gamma_u L_f}{\gamma_d L_0}. \tag{4.12}$$

Proof:
Denote by $n_i \geq 1$ the number of calls of the oracle at iteration $i \geq 0$. At each cycle of the internal loop we call the oracle twice for computing $\nabla f(y)$ and $\nabla f(T_L(y))$. Therefore,

$$L_{i+1} \;=\; \tfrac{1}{\gamma_d} \cdot L_i \cdot \gamma_u^{0.5 n_i - 1}.$$

Thus,

$$n_i \;=\; 2\left[1 + \tfrac{\ln \gamma_d}{\ln \gamma_u} + \tfrac{1}{\ln \gamma_u} \cdot \ln \tfrac{L_{i+1}}{L_i}\right].$$

Hence, we can compute

$$N_k \;=\; \sum_{i=0}^{k} n_i \;=\; 2\left[1 + \tfrac{\ln \gamma_d}{\ln \gamma_u}\right] \cdot (k+1) + \tfrac{1}{\ln \gamma_u} \cdot \ln \tfrac{L_{k+1}}{L_0}.$$

In remains to note that $L_{k+1} \overset{(4.11)}{\leq} \tfrac{\gamma_u}{\gamma_d} L_f$. $\square$

Thus, each iteration of (4.9) needs approximately two times more calls of oracle than one iteration of the Gradient Method:

$$\gamma_u \;=\; \gamma_d \;= 2 \quad \Rightarrow \quad N_k \;\leq\; 4(k+1) + \log_2 \tfrac{L_f}{L_0}, \quad L_k \;\leq L_f. \tag{4.13}$$

However, we will see that the rate of convergence of (4.9) is much higher.

Let us start from two auxiliary statements.

Lemma 6 *Assume $\mu_\Psi \geq \mu$. Then the sequences $\{x_k\}$, $\{A_k\}$ and $\{\psi_k\}$, generated by the method $\mathcal{A}(x_0, L_0, \mu)$, satisfy relations (4.4) for all $k \geq 0$, .*

Proof:
Indeed, in view of initial settings of (4.9), $A_0 = 0$ and $\psi_0^* = 0$. Hence, for $k = 0$, both relations (4.4) are trivial.

Assume now that relations $\mathcal{R}_k^1$, $\mathcal{R}_k^2$ are valid for some $k \geq 0$. In view of $\mathcal{R}_k^2$, for any $x \in E$ we have

$$\begin{aligned}
\psi_{k+1}(x) \;&\leq\; A_k \phi(x) + \tfrac{1}{2}\|x - x_0\|^2 + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)] \\
&\leq\; (A_k + a_{k+1})\phi(x) + \tfrac{1}{2}\|x - x_0\|^2,
\end{aligned}$$

and this is $\mathcal{R}^2_{k+1}$. Let us show that the relation $\mathcal{R}^1_{k+1}$ is also valid.

Indeed, in view of (4.3), function $\psi_k(x)$ is strongly convex with convexity parameter $1 + \mu A_k$. Hence, in view of $\mathcal{R}^1_k$, for any $x \in E$, we have

$$\psi_k(x) \;\geq\; \psi_k^* + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2 \;\geq\; A_k\phi(x_k) + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2. \tag{4.14}$$

Therefore

$$\begin{aligned}
\psi^*_{k+1} = \min_{x\in E} \;& \{\psi_k(x) + a_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)]\} \\
\overset{(4.14)}{\geq} \min_{x\in E} \;& \Big\{A_k\phi(x_k) + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2 + a_{k+1}[\phi(x_{k+1}) + \langle \phi'(x_{k+1}), x - x_{k+1}\rangle]\Big\} \\
\geq \min_{x\in E} \;& \{(A_k + a_{k+1})\phi(x_{k+1}) + A_k\langle \phi'(x_{k+1}), x_k - x_{k+1}\rangle \\
& + a_{k+1}\langle \phi'(x_{k+1}), x - x_{k+1}\rangle + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2\} \\
\overset{(4.9)}{=} \min_{x\in E} \;& \{A_{k+1}\phi(x_{k+1}) + \langle \phi'(x_{k+1}), A_{k+1}y_k - a_{k+1}v_k - A_k x_{k+1}\rangle \\
& + a_{k+1}\langle \phi'(x_{k+1}), x - x_{k+1}\rangle + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2\} \\
= \min_{x\in E} \;& \{A_{k+1}\phi(x_{k+1}) + A_{k+1}\langle \phi'(x_{k+1}), y_k - x_{k+1}\rangle \\
& + a_{k+1}\langle \phi'(x_{k+1}), x - v_k\rangle + \tfrac{1+\mu A_k}{2}\|x - v_k\|^2\}.
\end{aligned}$$

Thus, we have proved inequality

$$\psi^*_{k+1} \;\geq\; A_{k+1}\phi(x_{k+1}) + A_{k+1}\langle \phi'(x_{k+1}), y_k - x_{k+1}\rangle - \tfrac{a^2_{k+1}}{2(1+\mu A_k)}\|\phi'(x_{k+1})\|^2_*.$$

On the other hand, by termination criterion in (4.9), we have

$$\langle \phi'(x_{k+1}), y_k - x_{k+1}\rangle \;\geq\; \tfrac{1}{M_k}\|\phi'(x_{k+1})\|^2_*.$$

It remains to note that in (4.9) we choose a_{k+1} from the quadratic equation

$$A_{k+1} \;\equiv\; A_k + a_{k+1} \;=\; \tfrac{M_k a^2_{k+1}}{2(1+\mu A_k)}.$$

Thus, $\mathcal{R}^1_{k+1}$ is valid. □

Thus, in order to use inequality (4.5) for deriving the rate of convergence of method $\mathcal{A}(x_0, L_0, \mu)$, we need to estimate the rate of growth of the scaling coefficients $\{A_k\}_{k=0}^{\infty}$.

Lemma 7 *For any $\mu \geq 0$, the scaling coefficients grow as follows:*

$$A_k \;\geq\; \tfrac{k^2}{2\gamma_u L_f}, \quad k \geq 0. \tag{4.15}$$

For $\mu > 0$, the rate of growth is linear:

$$A_k \;\geq\; \tfrac{1}{\gamma_u L_f} \cdot \left[1 + \sqrt{\tfrac{\mu}{2\gamma_u L_f}}\right]^{2(k-1)}, \quad k \geq 1. \tag{4.16}$$

Proof:
Indeed, in view of equation $(*)$ in (4.9), we have:

$$A_{k+1} \;\leq\; A_{k+1}(1+\mu A_k) \;=\; \tfrac{M_k}{2}(A_{k+1}-A_k)^2 \;=\; \tfrac{M_k}{2}\left[A_{k+1}^{1/2}-A_k^{1/2}\right]^2\left[A_{k+1}^{1/2}+A_k^{1/2}\right]^2$$

$$\leq\; 2A_{k+1}M_k\left[A_{k+1}^{1/2}-A_k^{1/2}\right]^2 \overset{(4.11)}{\leq} 2A_{k+1}\gamma_u L_f\left[A_{k+1}^{1/2}-A_k^{1/2}\right]^2.$$

Thus, for any $k \geq 0$ we get $A_k^{1/2} \geq \frac{k}{\sqrt{2\gamma_u L_f}}$. If $\mu > 0$, then, by the same reasoning as above, we obtain

$$\mu A_k A_{k+1} \;<\; A_{k+1}(1+\mu A_k) \;\leq\; 2A_{k+1}\gamma_u L_f\left[A_{k+1}^{1/2}-A_k^{1/2}\right]^2.$$

Hence, $A_{k+1}^{1/2} \geq A_k^{1/2}\left[1+\sqrt{\frac{\mu}{2\gamma_u L_f}}\right]$. Since $A_1 = \frac{1}{M_0} \overset{(4.11)}{\geq} \frac{1}{\gamma_u L_f}$, we come to (4.16). □

Now we can summarize all our observations.

Theorem 6 *Let the gradient of function f be Lipschitz continuous with constant L_f. And let the parameter L_0 satisfy condition (3.2). Then the rate of convergence of the method $\mathcal{A}(x_0, L_0, 0)$ as applied to the problem (4.1) can be estimated as follows:*

$$\phi(x_k) - \phi(x^*) \;\leq \tfrac{\gamma_u L_f \|x^*-x_0\|^2}{k^2}, \quad k \geq 1. \tag{4.17}$$

If in addition the function Ψ is strongly convex, then the sequence $\{x_k\}_{k=1}^{\infty}$ generated by $\mathcal{A}(x_0, L_0, \mu_\Psi)$ satisfies both (4.17) and the following inequality:

$$\phi(x_k) - \phi(x^*) \;\leq \tfrac{\gamma_u L_f}{2}\|x^*-x_0\|^2 \cdot \left[1+\sqrt{\tfrac{\mu_\Psi}{2\gamma_u L_f}}\right]^{-2(k-1)}, \quad k \geq 1. \tag{4.18}$$

In the next section we will show how to apply this result in order to achieve some specific goals for different optimization problems.

5 Different minimization strategies

5.1 Strongly convex objective with known parameter

Consider the following convex constrained minimization problem:

$$\min_{x \in Q} \hat{f}(x), \tag{5.1}$$

where Q is a closed convex set, and $\hat{f}$ is a strongly convex function with Lipschitz continuous gradient. Assume the convexity parameter $\mu_{\hat{f}}$ to be known. Denote by $\sigma_Q(x)$ an indicator function of set Q:

$$\sigma_Q(x) \;=\; \begin{cases} 0, & x \in Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

We can solve the problem (5.1) by two different techniques.

1. Reallocating the prox-term in the objective. For $\mu \in (0, \mu_{\hat{f}}]$, define

$$f(x) \;=\; \hat{f}(x) - \tfrac{\mu}{2}\|x - x_0\|^2, \quad \Psi(x) \;=\; \sigma_Q(x) + \tfrac{\mu}{2}\|x - x_0\|^2. \tag{5.2}$$

Note that function f in (5.2) is convex and its gradient is Lipschitz continuous with $L_f = L_{\hat{f}} - \mu$. Moreover, the function $\Psi(x)$ is strongly convex with convexity parameter μ. On the other hand,

$$\phi(x) \;=\; f(x) + \Psi(x) \;=\; \hat{f}(x) + \sigma_Q(x).$$

Thus, the corresponding unconstrained minimization problem (4.1) coincides with constrained problem (5.1). Since all conditions of Theorem 6 are satisfied, the method $\mathcal{A}(x_0, L_0, \mu)$ has the following performance guarantees:

$$\hat{f}(x_k) - \hat{f}(x^*) \;\leq\; \frac{\gamma_u (L_{\hat{f}} - \mu)\|x^* - x_0\|^2}{2\left[1 + \sqrt{\frac{\mu}{2\gamma_u (L_{\hat{f}} - \mu)}}\right]^{2(k-1)}}, \quad k \geq 1. \tag{5.3}$$

This means that an ϵ-solution of problem (5.1) can be obtained by this technique in

$$O\left(\sqrt{\tfrac{L_{\hat{f}}}{\mu}} \cdot \ln \tfrac{1}{\epsilon}\right) \tag{5.4}$$

iterations. Note that the same problem can be solve also by the Gradient Method (3.3). However, in accordance to (3.15), its performance guarantee is much worse; it needs

$$O\left(\tfrac{L_{\hat{f}}}{\mu} \cdot \ln \tfrac{1}{\epsilon}\right)$$

iterations.

2. Restart. For problem (5.1), define the following components of composite objective function in (4.1):

$$f(x) \;=\; \hat{f}(x), \quad \Psi(x) \;=\; \sigma_Q(x). \tag{5.5}$$

Let us fix an upper bound $N \geq 1$ for the number of iterations in $\mathcal{A}$. Consider the following two-level process:

$$\begin{array}{l} \text{Choose } u_0 \in Q. \\ \\ \text{Compute } u_{k+1} \text{ as a result of } N \text{ iterations of } \mathcal{A}(u_k, L_0, 0),\ k \geq 0. \end{array} \tag{5.6}$$

In view of definition (5.5), we have

$$\hat{f}(u_{k+1}) - \hat{f}(x^*) \;\overset{(4.17)}{\leq}\; \tfrac{\gamma_u L_{\hat{f}}\|x^* - u_k\|^2}{N^2} \;\leq\; \tfrac{2\gamma_u L_{\hat{f}}[\hat{f}(u_k) - \hat{f}(x^*)]}{\mu_{\hat{f}} \cdot N^2}.$$

Thus, taking $N = 2\sqrt{\frac{\gamma_u L_{\hat{f}}}{\mu_{\hat{f}}}}$, we obtain

$$\hat{f}(u_{k+1}) - \hat{f}(x^*) \;\leq\; \tfrac{1}{2}[\hat{f}(u_{k+1}) - \hat{f}(x^*)].$$

Hence, the performance guarantees of this technique are of the same order as (5.4).

5.2 Approximating the first-order optimality conditions

In some applications, we are interested in finding a point with small residual of the system of the first-order optimality conditions. Since

$$D\phi(T_L(x))[u] \overset{(2.15)}{\geq} -\left(1+\tfrac{L_f}{L}\right)\cdot\|g_L(x)\|_* \overset{(2.12)}{\geq} -(L+L_f)\cdot\sqrt{\tfrac{\phi(x)-\phi(x^*)}{2L-L_f}} \quad \forall u\in\mathcal{F}(T_L(x)),\ \|u\|=1, \tag{5.7}$$

the upper bounds on this residual can be obtained from the estimates on the rate of convergence of method (4.9) in the form (4.17) or (4.18). However, in this case, the first inequality does not give a satisfactory result. Indeed, it can guarantee that the right-hand side of inequality (5.7) vanishes as $O(\frac{1}{k})$. This rate is typical for the Gradient Method (see (3.12)), and from accelerated version (4.9) we can expect much more. Let us show how we can achieve a better result.

Consider the following constrained optimization problem:

$$\min_{x\in Q} f(x), \tag{5.8}$$

where Q is a closed convex set, and f is a convex function with Lipschitz continuous gradient. Let us fix a tolerance parameter $\delta>0$ and a starting point $x_0\in Q$. Define

$$\Psi(x) \;=\; \sigma_Q(x)+\tfrac{\delta}{2}\|x-x_0\|^2.$$

Consider now the unconstrained minimization problem (4.1) with composite objective function $\phi(x)=f(x)+\Psi(x)$. Note that function Ψ is strongly convex with parameter $\mu_\Psi=\delta$. Hence, in view of Theorem 6, the method $\mathcal{A}(x_0,L_0,\delta)$ converges as follows:

$$\phi(x_k)-\phi(x^*) \;\leq\; \tfrac{\gamma_u L_f}{2}\|x^*-x_0\|^2\cdot\left[1+\sqrt{\tfrac{\delta}{2\gamma_u L_f}}\right]^{-2(k-1)}. \tag{5.9}$$

For simplicity, we can choose $\gamma_u=\gamma_d$ in order to have $L_k\leq L_f$ for all $k\geq 0$.

Let us compute now $T_k=\mathcal{G}(x_k,L_k).T$ and $M_k=\mathcal{G}(x_k,L_k).L$. Then

$$\phi(x_k)-\phi(x^*)\;\geq\;\phi(x_k)-\phi(T_k) \overset{(2.16)}{\geq} \tfrac{1}{2M_k}\|g_{M_k}(x_k)\|_*^2, \quad L_0\;\leq\;M_k\;\leq\;\gamma_u L_f,$$

and we obtain the following estimate:

$$\|g_{M_k}(x_k)\|_* \overset{(5.9)}{\leq} \gamma_u L_f\|x^*-x_0\|\cdot\left[1+\sqrt{\tfrac{\delta}{2\gamma_u L_f}}\right]^{1-k}. \tag{5.10}$$

In our case, the first-order optimality conditions (4.2) for computing $T_{M_k}(x_k)$ can be written as follows:

$$\nabla f(x_k)+\delta B(T_k-x_0)+\xi_k \;=\; g_{M_k}(x_k), \tag{5.11}$$

where $\xi_k\in\partial\sigma_Q(T_k)$. Note that for any $y\in Q$ we have

$$0 \;=\; \sigma_Q(y)\;\geq\sigma_Q(T_k)+\langle\xi_k,y-T_k\rangle\;=\;\langle\xi_k,y-T_k\rangle. \tag{5.12}$$

Hence, for any direction $u \in \mathcal{F}(T_k)$ with $\|u\| = 1$ we obtain

$$\begin{aligned}
\langle \nabla f(T_k), u \rangle &\overset{(2.10)}{\geq} \langle \nabla f(x_k), u \rangle - \tfrac{L_f}{M_k} \|g_{M_k}(x_k)\|_* \\
&\overset{(5.11)}{=} \langle g_{M_k}(x_k) - \delta B(T_k - x_0) - \xi_k, u \rangle - \tfrac{L_f}{M_k} \|g_{M_k}(x_k)\|_* \\
&\overset{(5.12)}{=} -\delta \cdot \|T_k - x_0\| - \left(1 + \tfrac{L_f}{M_k}\right) \cdot \|g_{M_k}(x_k)\|_*.
\end{aligned}$$

Assume now that the size of the set Q does not exceed R, and $\delta = \epsilon \cdot L_0$. Let us choose the number of iterations k from inequality

$$\left[1 + \sqrt{\tfrac{\epsilon L_0}{2\gamma_u L_f}}\right]^{1-k} \quad \leq \quad \epsilon.$$

Then the residual of the first-order optimality conditions satisfies the following inequality:

$$\langle \nabla f(T_k), u \rangle \quad \geq \quad -\epsilon \cdot R \cdot \left[L_0 + \gamma_u L_f \cdot \left(1 + \tfrac{L_f}{L_0}\right)\right], \quad u \in \mathcal{F}(T_k),\ \|u\| = 1. \tag{5.13}$$

For that, the required number of iterations k is at most of the order $O\left(\frac{1}{\sqrt{\epsilon}} \ln \frac{1}{\epsilon}\right)$.

5.3 Unknown parameter of strongly convex objective

In Section 5.1 we have discussed two efficient strategies for minimizing strongly convex function with known estimate of convexity parameter $\mu_{\hat{f}}$. However, usually this information is not available. We can easily get only an *upper* estimate for this value, for example, by inequality

$$\mu_{\hat{f}} \quad \leq \quad S_L(x) \ \leq \ L_{\hat{f}}, \quad x \in Q.$$

Let us show that such a bound can be also used for designing an efficient optimization strategy for strongly convex functions.

For problem (5.1), assume that we have some guess μ for the parameter $\mu_{\hat{f}}$ and a starting point $u_0 \in Q$. Denote $\phi_0(x) = \hat{f}(x) + \sigma_Q(x)$. Let us choose

$$x_0 = \mathcal{G}_{\phi_0}(u_0, L_0).T, \quad M_0 = \mathcal{G}_{\phi_0}(u_0, L_0).L, \quad S_0 = \mathcal{G}_{\phi_0}(u_0, L_0).S,$$

and minimize the composite objective (5.2) by method $\mathcal{A}(x_0, M_0, \mu)$, endowed by the following stopping criterion:

COMPUTE:	$v_k = \mathcal{G}_{\phi_0}(x_k, L_k).T, \quad M_k = \mathcal{G}_{\phi_0}(x_k, L_k).L.$	
STOP THE STAGE:	if (A): $\| g_{M_k}(x_k)[\phi_0] \|_* \leq \frac{1}{2} \| g_{M_0}(u_0)[\phi_0] \|_*,$	(5.14)
	or (B): $\frac{M_k}{A_k} \cdot \left(1 + \frac{S_0}{M_0}\right) \leq \frac{1}{4}\mu^2.$	

If the stage was terminated by Condition (A), then we call it *successful*. In this case, we run the next stage, taking v_k as a new starting point and keeping the estimate μ of the convexity parameter unchanged.

Suppose that the stage was terminated by Condition (B) (that is an *unsuccessful* stage). If μ would be a correct *lower* bound for the convexity parameter $\mu_{\hat{f}}$, then

$$\begin{aligned} \tfrac{1}{2M_k}\|g_{M_k}(x_k)[\phi_0]\|_*^2 &\overset{(2.16)}{\leq} \hat{f}(x_k) - \hat{f}(x^*) \overset{(4.5)}{\leq} \tfrac{1}{2A_k}\|x_0 - x^*\|^2 \\ &\overset{(2.21)}{\leq} \tfrac{1}{2A_k\mu^2} \cdot \left(1 + \tfrac{S_0}{M_0}\right) \cdot \|g_{M_0}(u_0)[\phi_0]\|_*^2. \end{aligned}$$

Hence, in view of Condition (B), in this case the stage must be terminated by Condition (A). Since this did not happen, we conclude that $\mu > \mu_{\hat{f}}$. Therefore, we redefine $\mu := \frac{1}{2}\mu$, and run again the stage keeping the old starting point x_0.

We are not going to present all details of the complexity analysis of the above strategy. It can be shown that, for generating an ϵ-solution of problem (5.1) with strongly convex objective, it needs

$$O\left(\kappa_{\hat{f}}^{1/2} \ln \kappa_{\hat{f}}\right) + O\left(\kappa_{\hat{f}}^{1/2} \ln \kappa_{\hat{f}} \cdot \ln \tfrac{\kappa_{\hat{f}}}{\epsilon}\right), \quad \kappa_{\hat{f}} \overset{\text{def}}{=} \tfrac{L_{\hat{f}}}{\mu_{\hat{f}}},$$

calls of oracle. The first term in this bound corresponds to the total amount of calls of oracle at all unsuccessful stages. The factor $\kappa_{\hat{f}}^{1/2} \ln \kappa_{\hat{f}}$ represents an upper bound on the length of any stage independently on the variant of its termination.

6 Computational experiments

We tested the above mentioned algorithms on a set of randomly generated *Sparse Least Squares* problems of the form

$$\text{Find } \phi^* = \min_{x \in R^n} \left[\phi(x) \overset{\text{def}}{=} \tfrac{1}{2}\|Ax - b\|_2^2 + \|x\|_1 \right], \tag{6.1}$$

where $A \equiv (a_1, \ldots, a_n)$ is an $m \times n$ dense matrix with $m < n$. All problems were generated with known optimal solutions, which can be obtained from the dual representation of the initial *primal* problem (6.1):

$$\begin{aligned} \min_{x \in R^n} \left[\tfrac{1}{2}\|Ax - b\|_2^2 + \|x\|_1 \right] &= \min_{x \in R^n} \max_{u \in R^m} \left[\langle u, b - Ax \rangle - \tfrac{1}{2}\|u\|_2^2 + \|x\|_1 \right] \\ &= \max_{u \in R^m} \min_{x \in R^n} \left[\langle b, u \rangle - \tfrac{1}{2}\|u\|_2^2 - \langle A^T u, x \rangle + \|x\|_1 \right] \qquad (6.2) \\ &= \max_{u \in R^m} \left[\langle b, u \rangle - \tfrac{1}{2}\|u\|_2^2 : \ \|A^T u\|_\infty \leq 1 \right]. \end{aligned}$$

Thus, the problem dual to (6.1) consists in finding a Euclidean projection of the vector $b \in R^m$ onto the dual polytop

$$\mathcal{D} \ = \ \{y \in R^m : \ \|A^T y\|_\infty \leq 1\}.$$

This interpretation explains the changing sparsity of the optimal solution $x^*(\tau)$ to the following parametric version of problem (6.1):

$$\min_{x \in R^n} \left[\phi_\tau(x) \overset{\text{def}}{=} \tfrac{1}{2}\|Ax - b\|_2^2 + \tau\|x\|_1 \right] \tag{6.3}$$

Indeed, for $\tau > 0$, we have

$$\phi_\tau(x) \;=\; \tau^2 \left[\tfrac{1}{2} \| A \tfrac{x}{\tau} - \tfrac{b}{\tau} \|_2^2 + \| \tfrac{x}{\tau} \|_1 \right].$$

Hence, in the dual problem, we project vector $\frac{b}{\tau}$ onto the polytop $\mathcal{D}$. The nonzero components of $x^*(\tau)$ correspond to the active facets of $\mathcal{D}$. Thus, for τ big enough, we have $\frac{b}{\tau} \in \operatorname{int} \mathcal{D}$, which means $x^*(\tau) = 0$. When τ decreases, we get $x^*(\tau)$ more and more dense. Finally, if all facets of $\mathcal{D}$ are in general position, we get in $x^*(\tau)$ exactly m nonzero components as $\tau \to 0$.

In our computational experiments, we compare three minimization methods. Two of them maintain recursively relations (4.4). This feature allows to classify them as the primal-dual methods. Indeed, denote

$$\phi_*(u) \;=\; \tfrac{1}{2} \|u\|_2^2 - \langle b, u \rangle.$$

As we have seen in (6.2),

$$\phi(x) + \phi_*(u) \;\geq\; 0 \quad \forall x \in R^n,\; u \in \mathcal{D}. \tag{6.4}$$

Moreover, the lower bound is achieved only at the optimal solutions of the primal and dual problems. For some sequence $\{z_i\}_{i=1}^\infty$, and a starting point $z_0 \in \operatorname{dom} \Psi$, relations (4.4) ensure

$$A_k \phi(x_k) \;\leq\; \min_{x \in R^n} \left\{ \sum_{i=1}^k a_i [f(z_i) + \langle \nabla f(z_i), x - z_i \rangle] + A_k \Psi(x) + \tfrac{1}{2} \|x - z_0\|^2 \right\}. \tag{6.5}$$

In our situation, $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $\Psi(x) = \|x\|_1$, and we choose $z_0 = 0$. Denote $u_i = b - Az_i$. Then $\nabla f(z_i) = -A^T u_i$, and therefore

$$f(z_i) - \langle \nabla f(z_i), z_i \rangle \;=\; \tfrac{1}{2}\|u_i\|^2 + \langle A^T u_i, z_i \rangle \;=\; \langle b, u_i \rangle - \tfrac{1}{2}\|u_i\|_2^2 \;=\; -\phi_*(u_i).$$

Denoting

$$\bar{u}_k \;=\; \tfrac{1}{A_k} \sum_{i=1}^k a_i u_i, \tag{6.6}$$

we obtain

$$\begin{aligned} A_k [\phi(x_k) + \phi_*(\bar{u}_k)] \;&\leq\; A_k \phi(x_k) + \sum_{i=1}^k a_i \phi_*(u_i) \\ &\overset{(6.5)}{\leq} \min_{x \in R^n} \left\{ \sum_{i=1}^k a_i \langle \nabla f(z_i), x \rangle + A_k \Psi(x) + \tfrac{1}{2}\|x\|^2 \right\} \overset{x=0}{\leq} 0. \end{aligned}$$

In view of (6.4), u_k cannot be feasible:

$$\phi_*(\bar{u}_k) \;\leq\; -\phi(x_k) \;\leq\; -\phi^* \;=\; \min_{u \in \mathcal{D}} \phi_*(u). \tag{6.7}$$

Let us measure the level of infeasibility of these points. Note that the minimum of optimization problem in (6.5) is achieved at $x = v_k$. Hence, the corresponding first-order optimality conditions ensure

$$\| - \sum_{i=1}^k a_i A^T u_i + B v_k \|_\infty \;\leq\; A_k.$$

Therefore, $|\langle a_i, \bar{u}_k \rangle| \leq 1 + \frac{1}{A_k} |(Bv_k)^{(i)}|$, $i = 1, \ldots, n$. Assume that the matrix B in (1.3) is diagonal:

$$B^{(i,j)} = \begin{cases} d_i, & i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Then, $|\langle a_i, \bar{u}_k \rangle| - 1 \leq \frac{d_i}{A_k} \cdot |v_k^{(i)}|$, and

$$\rho(\bar{u}_k) \stackrel{\text{def}}{=} \left[\sum_{i=1}^{n} \frac{1}{d_i} \cdot (\, |\langle a_i, \bar{u}_k \rangle| - 1 \,)_+^2 \right]^{1/2} \leq \frac{1}{A_k} \|v_k\| \stackrel{(4.6)}{\leq} \frac{2}{A_k} \|x^*\|, \qquad (6.8)$$

where $(\alpha)_+ = \max\{\alpha, 0\}$. Thus, we can use function $\rho(\cdot)$ as a dual infeasibility measure. In view of (6.7), it is a reasonable stopping criterion for our primal-dual methods.

For generating the random test problems, we apply the following strategy.

- Choose $m_* \leq m$, the number of nonzero components of the optimal solution x^* of problem (6.1), and parameter $\rho > 0$ responsible for the size of x^*.
- Generate randomly a matrix $B \in R^{m \times n}$ with elements uniformly distributed in the interval $[-1, 1]$.
- Generate randomly a vector $v^* \in R^m$ with elements uniformly distributed in $[0, 1]$. Define $y^* = v^* / \|v^*\|_2$.
- Sort the entries of vector $B^T y^*$ in the order of decrease of their absolute values. For the sake of notation, assume that this is a natural ordering.
- For $i = 1, \ldots, n$, define $a_i = \alpha_i b_i$ with $\alpha_i > 0$ chosen in accordance to the following rule:

$$\alpha_i = \begin{cases} \frac{1}{|\langle b_i, y^* \rangle|} & \text{for } i = 1, \ldots, m_*, \\ 1, & \text{if } |\langle b_i, y^* \rangle| \leq 0.1 \text{ and } i > m_*, \\ \frac{\xi_i}{|\langle b_i, y^* \rangle|}, & \text{otherwise,} \end{cases}$$

  where ξ_i are uniformly distributed in $[0, 1]$.
- For $i = 1, \ldots, n$, generate the components of the primal solution:

$$[x^*]^{(i)} = \begin{cases} \xi_i \cdot \text{sign}(\langle a_i, y^* \rangle), & \text{for } i \leq m_*, \\ 0, & \text{otherwise,} \end{cases}$$

  where ξ_i are uniformly distributed in $\left[0, \frac{\rho}{\sqrt{m_*}}\right]$.
- Define $b = y^* + Ax^*$.

Thus, the optimal value of the randomly generated problem (6.1) can be computed as

$$\phi^* = \frac{1}{2} \|y^*\|_2^2 + \|x^*\|_1.$$

In the first series of tests, we use this value in the termination criterion.

Let us look first at the results of minimization of two typical random problem instances. The first problem is relatively easy.

Problem 1: $n = 4000,\ m = 1000,\ m_* = 100,\ \rho = 1.$

	PG			DG			AC		
Gap	k	#Ax	SpeedUp	k	#Ax	SpeedUp	k	#Ax	SpeedUp
1	1	4	0.21%	1	4	0.85%	1	4	0.14%
2^{-1}	3	8	0.20%	3	12	0.81%	4	28	1.24%
2^{-2}	10	29	0.24%	8	38	0.89%	8	60	2.47%
2^{-3}	28	83	0.32%	25	123	1.17%	14	108	4.06%
2^{-4}	159	476	0.88%	156	777	3.45%	40	316	17.50%
2^{-5}	557	1670	1.53%	565	2824	6.21%	74	588	29.47%
2^{-6}	954	2862	1.31%	941	4702	5.17%	98	780	25.79%
2^{-7}	1255	3765	0.86%	1257	6282	3.45%	118	940	18.62%
2^{-8}	1430	4291	0.49%	1466	7328	2.01%	138	1096	12.73%
2^{-9}	1547	4641	0.26%	1613	8080	2.13%	156	1240	8.19%
2^{-10}	1640	4920	0.14%	1743	8713	0.61%	173	1380	4.97%
2^{-11}	1722	5167	0.07%	1849	9243	0.33%	188	1500	3.01%
2^{-12}	1788	5364	0.04%	1935	9672	0.17%	202	1608	1.67%
2^{-13}	1847	5539	0.02%	2003	10013	0.09%	216	1720	0.96%
2^{-14}	1898	5693	0.01%	2061	10303	0.05%	230	1836	0.55%
2^{-15}	1944	5831	0.01%	2113	10563	0.05%	248	1968	0.31%
2^{-16}	1987	5961	0.00%	2164	10817	0.03%	265	2112	0.19%
2^{-17}	2029	6085	0.00%	2217	11083	0.02%	279	2224	0.10%
2^{-18}	2072	6215	0.00%	2272	11357	0.01%	305	2432	0.06%
2^{-19}	2120	6359	0.00%	2331	11652	0.00%	314	2504	0.03%
2^{-20}	2165	6495	0.00%	2448	12238	0.00%	319	2544	0.02%

In this table, the column Gap shows the relative decrease of the initial residual. In the rest part of the table, we can see the computational results of three methods:

- Primal gradient method (3.3) abbreviated as PG.
- Dual version of the gradient method (4.7) abbreviated as DG.
- Accelerated gradient method (4.9) abbreviated as AC.

In all methods, we use the following values of the parameters:

$$\gamma_u = \gamma_d = 2, \quad x_0 = 0, \quad L_0 = \max_{1 \le i \le n} \|a_i\|^2 \ \le \ L_f, \quad \mu = 0.$$

Let us explain the remaining columns of this table. For each method, the column k shows the number of iterations necessary for reaching the corresponding reduction of the initial gap in the function value. Column Ax shows the necessary number of matrix-vector multiplications. Note that for computing the value $f(x)$ we need one multiplication. If in addition, we need to compute the gradient, we need one more multiplication. For example, in accordance to the estimate (4.13), each iteration of (4.9) needs four computations of the pair function/gradient. Hence, in this method we can expect eight matrix-vector multiplications per iteration. For the Gradient Method (3.3), we need in

average two calls of oracle. However, one of them is done in the "line-search" procedure (3.1) and it requires only the function value. Hence, in this case we expect to have three matrix-vector multiplications per iteration. In the above table, we can observe a remarkable accuracy of our predictions. Finally, the column SpeedUp represents the absolute accuracy of current approximate solution in percents to the worst-case estimate given by the corresponding rate of convergence. Since the exact L_f is unknown, we use L_0 instead.

We can see that all methods usually significantly outperform the theoretically predicted rate of convergence. However, for all of them, there are some parts of trajectory where the worst-case predictions are quite accurate. This is even more evident from our second table, which corresponds to a more difficult problem instance.

Problem 2: $n = 5000,\ m = 500,\ m_* = 100,\ \rho = 1.$

	PG			DG			AC		
Gap	k	#Ax	SpeedUp	k	#Ax	SpeedUp	k	#Ax	SpeedUp
1	1	4	0.24%	1	4	0.96%	1	4	0.16%
2^{-1}	2	6	0.20%	2	8	0.81%	3	24	0.92%
2^{-2}	5	17	0.21%	5	24	0.81%	5	40	1.49%
2^{-3}	11	33	0.19%	11	45	0.77%	8	64	1.83%
2^{-4}	38	113	0.30%	38	190	1.21%	19	148	5.45%
2^{-5}	234	703	0.91%	238	1189	3.69%	52	416	20.67%
2^{-6}	1027	3081	1.98%	1026	5128	7.89%	106	848	43.08%
2^{-7}	2402	7206	2.31%	2387	11933	9.17%	160	1280	48.70%
2^{-8}	3681	11043	1.77%	3664	18318	7.05%	204	1628	39.54%
2^{-9}	4677	14030	1.12%	4664	23318	4.49%	245	1956	28.60%
2^{-10}	5410	16230	0.65%	5392	26958	2.61%	288	2300	19.89%
2^{-11}	5938	17815	0.36%	5879	29393	1.41%	330	2636	13.06%
2^{-12}	6335	19006	0.19%	6218	31088	0.77%	370	2956	8.20%
2^{-13}	6637	19911	0.10%	6471	32353	0.41%	402	3212	4.77%
2^{-14}	6859	20577	0.05%	6670	33348	0.21%	429	3424	2.71%
2^{-15}	7021	21062	0.03%	6835	34173	0.13%	453	3616	1.49%
2^{-16}	7161	21483	0.01%	6978	34888	0.05%	471	3764	0.83%
2^{-17}	7281	21842	0.01%	7108	35539	0.05%	485	3872	0.42%
2^{-18}	7372	22115	0.00%	7225	36123	0.03%	509	4068	0.24%
2^{-19}	7438	22313	0.00%	7335	36673	0.02%	525	4192	0.12%
2^{-20}	7492	22474	0.00%	7433	37163	0.01%	547	4372	0.07%

In this table, we can see that the Primal Gradient Method still significantly outperforms the theoretical predictions. This is not too surprising since it can, for example, automatically accelerate on strongly convex functions (see Theorem 5). All other methods require in this case some explicit changes in their schemes.

However, despite to all these discrepancies, the main conclusion of our theoretical analysis seems to be confirmed: the accelerated scheme (4.9) significantly outperforms the primal and dual variants of the Gradient Method.

In the second series of tests, we studied the abilities of the primal-dual schemes (4.7) and (4.9) in decreasing the infeasibility measure $\rho(\cdot)$ (see (6.8)). This problem, at least for the Dual Gradient Method (4.7), appears to be much harder than the primal minimization

problem (6.1). Let us look at the following results.

Problem 3: $n = 500,\ m = 50,\ m_* = 25,\ \rho = 1.$

	DG				AC			
Gap	k	#Ax	$\Delta\phi$	SpeedUp	k	#Ax	$\Delta\phi$	SpeedUp
1	2	8	$2.5 \cdot 10^{0}$	8.26%	2	16	$3.6 \cdot 10^{0}$	2.80%
2^{-1}	5	25	$1.4 \cdot 10^{0}$	9.35%	7	56	$8.8 \cdot 10^{-1}$	15.55%
2^{-2}	13	64	$6.0 \cdot 10^{-1}$	13.17%	11	88	$5.3 \cdot 10^{-1}$	20.96%
2^{-3}	26	130	$3.9 \cdot 10^{-1}$	12.69%	15	120	$4.4 \cdot 10^{-1}$	19.59%
2^{-4}	48	239	$2.7 \cdot 10^{-1}$	12.32%	21	164	$3.1 \cdot 10^{-1}$	19.21%
2^{-5}	103	514	$1.6 \cdot 10^{-1}$	13.28%	35	276	$1.8 \cdot 10^{-1}$	25.83%
2^{-6}	243	1212	$8.3 \cdot 10^{-2}$	15.64%	54	432	$1.0 \cdot 10^{-1}$	31.75%
2^{-7}	804	4019	$3.0 \cdot 10^{-2}$	25.93%	86	688	$4.6 \cdot 10^{-2}$	39.89%
2^{-8}	1637	8183	$6.3 \cdot 10^{-3}$	26.41%	122	976	$1.8 \cdot 10^{-2}$	40.22%
2^{-9}	3298	16488	$4.6 \cdot 10^{-4}$	26.6%	169	1348	$5.3 \cdot 10^{-3}$	38.58%
2^{-10}	4837	24176	$1.8 \cdot 10^{-7}$	19.33%	224	1788	$7.7 \cdot 10^{-4}$	34.28%
2^{-11}	4942	24702	$1.2 \cdot 10^{-14}$	9.97%	301	2404	$8.0 \cdot 10^{-5}$	30.88%
2^{-12}	5149	25734	$-1.3 \cdot 10^{-15}$	5.16%	419	3352	$2.7 \cdot 10^{-5}$	29.95%
2^{-13}	5790	28944	$-1.3 \cdot 10^{-15}$	2.92%	584	4668	$5.3 \cdot 10^{-6}$	29.11%
2^{-14}	6474	32364	0.0	2.67%	649	5188	$4.1 \cdot 10^{-7}$	29.48%

In this table we can see the computational cost for decreasing the initial value of ρ in $2^{14} \approx 10^4$ times. Note that both methods require more iterations than for Problem 1, which was solved up to accuracy in the objective function of the order $2^{-20} \approx 10^{-6}$. Moreover, for reaching the required level of ρ, method (4.7) has to decrease the residual in the objective up to machine precision, and the norm of gradient mapping up to 10^{-12}. The accelerated scheme is more balanced: the final residual in ϕ is of the order 10^{-6}, and the norm of the gradient mapping was decreased only up to $1.3 \cdot 10^{-3}$.

Let us look at a bigger problem.

Problem 4: $n = 1000,\ m = 100,\ m_* = 50,\ \rho = 1.$

	DG				AC			
Gap	k	#Ax	$\Delta\phi$	SpeedUp	k	#Ax	$\Delta\phi$	SpeedUp
1	2	8	$3.7 \cdot 10^{0}$	6.41%	2	12	$4.2 \cdot 10^{0}$	1.99%
2^{-1}	5	24	$2.0 \cdot 10^{0}$	7.75%	7	56	$1.4 \cdot 10^{0}$	11.71%
2^{-2}	15	74	$1.0 \cdot 10^{0}$	11.56%	12	96	$8.7 \cdot 10^{-1}$	15.49%
2^{-3}	37	183	$6.9 \cdot 10^{-1}$	14.73%	17	132	$6.8 \cdot 10^{-1}$	16.66%
2^{-4}	83	414	$4.5 \cdot 10^{-1}$	16.49%	26	208	$4.7 \cdot 10^{-1}$	20.43%
2^{-5}	198	989	$2.4 \cdot 10^{-1}$	19.79%	42	336	$2.5 \cdot 10^{-1}$	26.76%
2^{-6}	445	2224	$7.8 \cdot 10^{-2}$	22.28%	65	520	$1.0 \cdot 10^{-1}$	32.41%
2^{-7}	1328	6639	$2.2 \cdot 10^{-2}$	33.25%	91	724	$3.6 \cdot 10^{-2}$	31.50%
2^{-8}	2675	13373	$4.1 \cdot 10^{-3}$	33.48%	125	996	$1.1 \cdot 10^{-2}$	30.07%
2^{-9}	4508	22535	$5.6 \cdot 10^{-5}$	28.22%	176	1404	$2.6 \cdot 10^{-3}$	27.85%
2^{-10}	4702	23503	$2.7 \cdot 10^{-10}$	14.7%	240	1916	$4.4 \cdot 10^{-4}$	26.08%
2^{-11}	4869	24334	$-2.2 \cdot 10^{-15}$	7.61%	328	2620	$7.7 \cdot 10^{-5}$	26.08%
2^{-12}	6236	31175	$-2.2 \cdot 10^{-15}$	4.88%	465	3716	$6.5 \cdot 10^{-6}$	26.20%
2^{-13}	12828	64136	$-2.2 \cdot 10^{-15}$	5.02%	638	5096	$2.4 \cdot 10^{-6}$	24.62%
2^{-14}	16354	81766	$-4.4 \cdot 10^{-15}$	5.24%	704	5628	$7.8 \cdot 10^{-7}$	24.62%

As compared with Problem 3, in Problem 4 the sizes are doubled. This makes almost no difference for the accelerated scheme, but for the Dual Gradient Method, the computational expenses grow substantially. The further increase of dimension makes the latter scheme impractical. Let us look how these methods work at Problem 1 with $\rho(\cdot)$ being a termination criterion.

Problem 1a: $n = 4000,\ m = 1000,\ m_* = 100,\ \rho = 1.$

	DG				AC			
Gap	k	#Ax	$\Delta\phi$	SpeedUp	k	#Ax	$\Delta\phi$	SpeedUp
1	2	8	$2.3 \cdot 10^{1}$	2.88%	2	12	$2.4 \cdot 10^{1}$	0.99%
2^{-1}	5	24	$1.2 \cdot 10^{1}$	3.44%	8	60	$8.1 \cdot 10^{0}$	7.02%
2^{-2}	17	83	$5.8 \cdot 10^{0}$	6.00%	13	100	$4.6 \cdot 10^{0}$	10.12%
2^{-3}	44	219	$3.5 \cdot 10^{0}$	7.67%	20	160	$3.5 \cdot 10^{0}$	11.20%
2^{-4}	100	497	$2.7 \cdot 10^{0}$	8.94%	28	220	$2.9 \cdot 10^{0}$	12.10%
2^{-5}	234	1168	$1.9 \cdot 10^{0}$	10.51%	44	348	$2.1 \cdot 10^{0}$	14.79%
2^{-6}	631	3153	$1.0 \cdot 10^{0}$	14.18%	78	620	$1.0 \cdot 10^{0}$	23.46%
2^{-7}	1914	9568	$1.0 \cdot 10^{-2}$	21.50%	117	932	$2.9 \cdot 10^{-1}$	26.44%
2^{-8}	3704	18514	$4.6 \cdot 10^{-7}$	20.77%	157	1252	$6.8 \cdot 10^{-2}$	23.88%
2^{-9}	3731	18678	$1.4 \cdot 10^{-14}$	15.77%	212	1688	$5.3 \cdot 10^{-3}$	21.63%
2^{-10}	Line	search	failure ...		287	2288	$2.0 \cdot 10^{-4}$	19.87%
2^{-11}					391	3120	$2.5 \cdot 10^{-5}$	18.43%
2^{-12}					522	4168	$7.0 \cdot 10^{-6}$	16.48%
2^{-13}					693	5536	$4.5 \cdot 10^{-7}$	14.40%
2^{-14}					745	5948	$3.8 \cdot 10^{-7}$	13.76%

The reason of the failure of the Dual Gradient Method is quite interesting. In the end, it generates the points with very small residual in the value of the objective function. Therefore, the termination criterion in the gradient iteration (3.1) cannot work properly due to the rounding errors. In the accelerated scheme (4.9), this does not happen since the decrease of the objective function and the dual infeasibility measure is much more balanced. In some sense, this situation is natural. We have seen that at the current test problems all methods converge faster in the end. On the other hand, the rate of convergence of the dual variables $\bar{u}_k$ is limited by the rate of growth of coefficients a_i in the representation (6.6). For the Dual Gradient Method, these coefficients are almost constant. For the accelerated scheme, they grow proportionally to the iteration counter.

We hope that the above numerical examples clearly demonstrate the advantages of the accelerated gradient method (4.9) with the adjustable line search strategy. It is interesting to check numerically how this method works in other situations. Of course, the first candidates to try are different applications of the smoothing technique [9]. However, even for the Sparse Least Squares problem (6.1) there are many potential improvements. Let us discuss one of them.

Note that we treated the problem (6.1) by a quite general model (2.1) ignoring the important fact that the function f is *quadratic*. The characteristic property of quadratic functions is that they have a constant second derivative. Hence, it is natural to select the operator B in metric (1.3) taking into account the structure of the Hessian of function f.

Let us define $B = \mathrm{diag}\,(A^T A) \equiv \mathrm{diag}\,(\nabla^2 f(x))$. Then

$$\|e_i\|^2 \;=\; \langle Be_i, e_i\rangle \;=\; \|a_i\|_2^2 \;=\; \|Ae_i\|_2^2 \;=\; \langle \nabla^2 f(x)e_i, e_i\rangle, \quad i = 1, \ldots, n,$$

where e_i is a coordinate vector in R^n. Therefore,

$$L_0 \stackrel{\rm def}{=} 1 \leq L_f \equiv \max_{\|u\|=1} \langle \nabla^2 f(x)u, u \rangle \leq n.$$

Thus, in this metric, we have very good lower and upper bounds for the Lipschitz constant L_f. Let us look at the corresponding computational results. We solve the Problem 1 (with $n = 4000$, $m = 1000$, and $\rho = 1$) up to accuracy GAP $= 2^{-20}$ for different sizes m_* of the support of the optimal vector, which are gradually increased from 100 to 1000.

Problem 1b.

	PG		AC	
m_*	K	#Ax	K	#Ax
100	42	127	58	472
200	53	160	61	496
300	69	208	70	568
400	95	286	77	624
500	132	397	84	680
600	214	642	108	872
700	330	993	139	1120
800	504	1513	158	1272
900	1149	3447	196	1576
1000	2876	8630	283	2272

Recall that the first line of this table corresponds to the previously discussed version of Problem 1. For the reader convenience, in the next table we repeat the final results on the latter problem again, adding the computational results for $m_* = 1000$, both with no diagonal scaling.

	PG		AC	
m_*	K	#Ax	K	#Ax
100	2165	6495	319	2544
1000	42509	127528	879	7028

Thus, for $m_* = 100$, the diagonal scaling makes Problem 1 *very easy*. For easy problems, the simple and cheap methods have definite advantage with respect to more complicated strategies. When m_* increases, the scaled problems become more and more difficult. Finally, we can see again the superiority of the accelerated scheme. Needless to say that at this moment of time, we have no plausible explanation for this phenomena.

Our last computational results clearly show that an appropriate complexity analysis of the Sparse Least Squares problem remains a challenging topic for the future research.

References

[1] S.Chen, D.Donoho, and M.Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computation*, **20** 33-61 (1998)

[2] J.Claerbout and F.Muir. Robust modelling of eratic data. *Geophysics*, **38** 826-844 (1973)

[3] M.Figueiredo, R.Novak, and S.Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. Submitted to publication.

[4] S.-J. Kim, K.Koh, M.Lustig, S.Boyd, and D.Gorinevsky. A method for large-scale l_1-regularized least-squares problems with applications in signal processing and statistics. Research Report, Stanford University, March 20, 2007.

[5] S.Levy and P.Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, **46** 1235-1243 (1981)

[6] A.Miller. Subset selection in regression. *Chapman and Hall*, London (2002)

[7] A.Nemirovsky and D.Yudin. Informational complexity and efficient methods for solution of convex extremal problems. J.Wiley & Sons, New-York (1983)

[8] Yu. Nesterov. Introductory Lectures on Convex Optimization. *Kluwer*, Boston (2004)

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

[10] Yu. Nesterov. Rounding of convex sets and efficient gradient methods for linear programming problems. Accepted by *Optimization Methods and Software* (2007).

[11] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. Accepted by *Mathematical Programming*. DOI 10.1007/s10107-006-0089-x.

[12] Yu. Nesterov and A. Nemirovskii. Interior point polynomial methods in convex programming: Theory and Applications. *SIAM*, Philadelphia (1994)

[13] J.Ortega and W.Rheinboldt. Iterative solution of nonlinear equations in several variables. *Academic Press*, New York (1970)

[14] F.Santosa and W.Symes. Linea inversion of band-limited reflection histograms. *SIAM Journal of Scientific and Statistical Computing*, **7** 1307-1330 (1986)

[15] H.Taylor, S.Bank, and J.McCoy. Deconvolution with the l_1 norm. *Geophysics*, **44** 39-52 (1979)

[16] R.Tibshirani. Regression shrinkage and selection via the lasso. *Journal Royal Statistical Society B*, **58** 267-288 (1996)

[17] J.Tropp. Just relax: convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory*, **51** 1030-1051 (2006)

[18] S.Wright. Solving l_1-regularized regression problems. Talk at International Conference "*Combinatorics and Optimization*", Waterloo, June 2007.