On Some Methods for Strongly Convex Optimization Problems with One Functional Constraint

Fedor S. Stonyakin[1](✉), Mohammad S. Alkousa[2], Alexander A. Titov[2], and Victoria V. Piskunova[1]

[1] V.I. Vernadsky Crimean Federal University, Simferopol, Russia
fedyor@mail.ru, viktoryapiskunova@yandex.ru

[2] Moscow Institute of Physics and Technology, Moscow, Russia
{mohammad.alkousa,a.a.titov}@phystech.edu

Abstract. We consider the classical optimization problem of minimizing a strongly convex, non-smooth, Lipschitz-continuous function with one Lipschitz-continuous constraint. We develop the approach in [10] and propose two methods for the considered problem with adaptive stopping rules. The main idea of the methods is using the dichotomy method and solving an auxiliary one-dimensional problem at each iteration. Theoretical estimates for the proposed methods are obtained. Partially, for smooth functions, we prove the linear rate of convergence of the methods. We also consider theoretical estimates in the case of non-smooth functions. The results for some examples of numerical experiments illustrating the advantages of the proposed methods and the comparison with some adaptive optimal method for non-smooth strongly convex functions are also given.

Keywords: Optimization with functional constraint · Adaptive method · Lipschitz-continuous function · Lipschitz-continuous gradient · Strongly convex objective function · Dichotomy method

1 Introduction

The optimization of non-smooth functions with constraints attracts wide interest in large-scale optimization and its applications [4,14]. There are a lot of methods of solving such kind of optimization problems. Some examples of these methods, to name but a few, are: bundle-level method [13], penalty method

The authors are very grateful to Alexander V. Gasnikov and Anastasiya S. Ivanova for fruitful discussions. The research of F. Stonyakin in Algorithm 1, Theorem 2 and Lemma 2 was supported by Russian Foundation for Basic Research according to the project 18-29-03071 mk. The research of F. Stonyakin in Subsects. 3.4 and 3.5 was supported by Russian Science Foundation grant 18-71-10044.

M. Khachay et al. (Eds.): MOTOR 2019, LNCS 11548, pp. 82–96, 2019.
https://doi.org/10.1007/978-3-030-22629-9_7

[15] and Lagrange multipliers method [5]. Recently in [2], some adaptive Mirror Descent methods were proposed for optimization problems of convex and strongly convex functions with non-smooth constraints

$$\min\{f(x) : \quad x \in Q \subset E, \quad g(x) \leq 0\}, \tag{1}$$

where Q is a convex and compact subset of a finite-dimensional real vector space E, $f : Q \to \mathbb{R}$ and $g : E \to \mathbb{R}$ are convex Lipschitz-continuous functions. In the case of several strongly convex non-smooth constraints, we consider one max-type constraint which is also strongly convex.

Methods in [2] are optimal from the point of view of lower oracle bounds and guarantee achieving acceptable precision ε with complexity $O\left(\varepsilon^{-1}\right)$ for strongly convex, Lipschitz-continuous objective f and convex Lipschitz-continuous constraint g.

In this paper, we develop the approach in [10] and propose an alternative approach for the problem (1) with a strongly convex Lipschitz-continuous objective f and a convex Lipschitz-continuous constraint g. Our approach is based on the transition to a strongly convex dual problem. In this case, the dual function depends on one dual variable $\lambda \geq 0$. When the Slater conditions for the problem (1) hold, all possible values of the dual variable are limited to a certain segment. This allows us to apply the dichotomy method similarly to [10] to search for the value of the dual variable λ, which is close to the appropriate λ_*, for which

$$\lambda_* \cdot g(x(\lambda_*)) = 0. \tag{2}$$

We propose two algorithms with adaptive stopping criterion that meet the necessary condition (2) in the general situation $\lambda_* \geq 0$ (Algorithm 1), as well as under the stronger assumption of the existence of $\lambda_* > 0$ (Algorithm 2). Partially, the last condition holds for the economic problem considered in [10].

It turns out that, with the possibility of a relatively quick solution of auxiliary problems, due to the proposed adaptive stopping criterion, Algorithms 1 and 2 may work faster than the optimal schemes in [2]. In proposed Algorithms 1 and 2 strong convexity of g is not required, and there is also no need to know the value of the strong convexity parameter of f.

The paper consists of an Introduction and four main sections. In Sect. 2 we consider the problem statement and some basic information concerning the necessary conditions of the extremum. In Sect. 3 we describe two main algorithms and give some estimates of the rate of convergence for them. Section 4 is devoted to basic information for optimal Mirror Descent Algorithms in the class of non-smooth strongly convex functions [2]. In Sect. 5 we make a comparison between the proposed algorithms and Mirror Descent Algorithm [2].

Thus, in the paper, we propose two methods for solving the problem (1) with the following types of assumptions:

$$|f(x) - f(y)| \leqslant M_f ||x - y||_2, \quad |g(x) - g(y)| \leqslant M_g ||x - y||_2 \tag{3}$$

or

$$||\nabla f(x) - \nabla f(y)||_2 \leqslant L_f ||x - y||_2, \quad ||\nabla g(x) - \nabla g(y)||_2 \leqslant L_g ||x - y||_2 \tag{4}$$

for all $x, y \in Q$, and for some real positive numbers M_f, M_g, L_f, L_g.

The contributions of this paper can be summarized as follows.

- With assumptions (4), the proposed methods have complexity

$$O\left(\log_2^2 \frac{1}{\varepsilon}\right), \tag{5}$$

  i.e. the linear rate of convergence. Note that we assume the strong convexity for the objective f only. The functional constraint g may not be strongly convex.
- With assumptions (3) we obtain complexity $O\left(\frac{1}{\varepsilon^2}\log_2\frac{1}{\varepsilon}\right)$, which is generally not optimal. However, due to the adaptivity of Algorithms 1 and 2, these methods can work faster than the optimal ones in [2] (see Sect. 5 below). Note that, unlike ([2], Subection 3.2), we require the strong convexity only of the objective functional f. In this case, the functional g, in general, may not be strongly convex.
- Also, a class of non-smooth functionals is considered, for which Algorithms 1 and 2 have complexity (5) (see Subsect. 3.4 below).

2 Problem Statement

Let $(E, ||\cdot||_2)$ be a normed finite-dimensional vector space with inner product $\langle\cdot,\cdot\rangle$ and norm $||x||_2 = \sqrt{\langle x, x\rangle}$. In this paper we consider the following optimization problem

$$f(x) \to \min_{\substack{g(x)\leqslant 0\\ x\in Q}}, \tag{6}$$

where f is a μ_f-strongly convex function with respect to the 2-norm, i.e.

$$f(\alpha x + (1-\alpha)y)) \leq \alpha f(x) + (1-\alpha)f(y) - \alpha(1-\alpha)\frac{\mu_f}{2}\|x-y\|_2^2$$

for $\alpha \in [0,1]$ and for all $x, y \in Q$. Assume that f and g are Lipschitz-continuous:

$$|f(y) - f(x)| \leqslant M_f\|y-x\|_2, \;\; \forall x, y \in Q,$$

$$|g(y) - g(x)| \leqslant M_g\|y-x\|_2, \;\; \forall x, y \in Q.$$

Let us introduce a dual factor $\lambda \geqslant 0$ and consider the dual problem to (6).

$$\min_{\substack{g(x)\leqslant 0\\ x\in Q}} f(x) = \min_{x\in Q}\left\{f(x) + \max_{\lambda\geqslant 0}(\lambda g(x))\right\} = \max_{\lambda\geqslant 0}\left\{\underbrace{\min_{x\in Q}\left(f(x) + \lambda g(x)\right)}_{=\varphi(\lambda)}\right\}.$$

Then the dual problem to the problem (6) is:

$$\varphi(\lambda) = f(x(\lambda)) + \lambda g(x(\lambda)) \to \max_{\lambda\geqslant 0}, \tag{7}$$

where

$$x(\lambda) = \arg\min_{x \in Q} \left\{ f(x) + \lambda g(x) \right\}. \tag{8}$$

Let us mention the following important well-known Demyanov-Danskin-Rubinov Theorem, see [7,8].

Theorem 1. *Let* $\varphi(\lambda) = \min\limits_{x \in X} F(x,\, \lambda)$ *for all* $\lambda \geqslant 0$*, where* $F(x,\, \lambda)$ *is a smooth convex function with respect to* λ *and* $x(\lambda)$ *is the only maximum point. Then*

$$\varphi'(\lambda) = F'_\lambda(x(\lambda),\, \lambda).$$

For the problem (7) Theorem 1 means that:

$$\varphi'(\lambda) = g(x(\lambda)). \tag{9}$$

Let λ^* be a solution of the dual problem (7). Then, according to the necessary condition of the extremum, the following equality must be satisfied for λ^*:

$$\lambda^* g(x(\lambda^*)) = 0,\ \lambda^* \geqslant 0,$$

which, by using (9), can be modified as follows:

$$\lambda^* \varphi'(\lambda^*) = 0,\ \lambda^* \geqslant 0. \tag{10}$$

3 Algorithms and Estimates of the Accuracy of Solutions and the Rate of Convergence

To solve the above-mentioned optimization problem (6), we proposed two algorithms. The main idea of the proposed algorithms is using the dichotomy method to solve the dual problem and solving an auxiliary one-dimensional problem at each iteration of the algorithms. Note that stopping criteria are the only difference between these algorithms.

Algorithm 1

```
Require: convex function $f$; initial localization interval $\left[\lambda^0_{min}, \lambda^0_{max}\right]$ of the dual variable; accuracy $\delta$ for auxiliary problems; accuracy $\varepsilon$.
$N := 0$
repeat
    $\lambda^N := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
    $x_\delta(\lambda^N) = \arg\min_{x \in Q}\{f(x) + \lambda^N g(x)\}$;
    $\varphi'(\lambda^N) = g(x_\delta(\lambda^N))$;
    if $\varphi'(\lambda^N) < 0$ then $\lambda^{N+1}_{max} := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
    if $\varphi'(\lambda^N) > 0$ then $\lambda^{N+1}_{min} := \frac{\lambda^N_{min} + \lambda^N_{max}}{2}$;
    $N := N + 1$;
until $\lambda^N |g(x_\delta(\lambda^N))| \le \varepsilon$.
Ensure: $\lambda^N$, with $\lambda^N |g(x_\delta(\lambda^N))| \le \varepsilon$; $x_\delta(\lambda^N)$.
```

Algorithm 2

```
Require: convex function $f$; initial localization interval $\left[\lambda_{min}^0, \lambda_{max}^0\right]$ of the dual variable; accuracy $\delta$ for auxiliary problems; accuracy $\varepsilon$.
$N := 0$
repeat
    $\lambda^N := \frac{\lambda_{min}^N + \lambda_{max}^N}{2}$;
    $x_\delta(\lambda^N) = \arg\min_{x \in Q}\{f(x) + \lambda^N g(x)\}$;
    $\varphi'(\lambda^N) = g(x_\delta(\lambda^N))$;
    if $\varphi'(\lambda^N) < 0$ then $\lambda_{max}^{N+1} := \frac{\lambda_{min}^N + \lambda_{max}^N}{2}$;
    if $\varphi'(\lambda^N) > 0$ then $\lambda_{min}^{N+1} := \frac{\lambda_{min}^N + \lambda_{max}^N}{2}$;
    $N := N + 1$;
until $|g(x_\delta(\lambda^N))| \leq \varepsilon$.
Ensure: $\lambda^N$, with $|g(x_\delta(\lambda^N))| \leq \varepsilon$; $x_\delta(\lambda^N)$.
```

Remark 1. Note that the stopping criterion of Algorithm 1 is necessarily reached due to the assumption that there exists such $k \in \mathbb{N}$, $\lambda^k = 0$. However, we need an additional assumption to guarantee that the Algorithm 2 stops. Suppose there exists a point $\overline{x} \in Q$, such that $g'(\overline{x}) = 0$.

3.1 Slater Condition

In order to use the dichotomy method and solve the dual problem, it is necessary to compactify the dual variable. So, the initial interval of the localization of the dual variable must be determined. As the dual variable reflects namely the inequality constraint, we can take zero as the lower bound, that means

$$\lambda_{min} = 0.$$

To determine the upper bound, we need to use the Slater condition.

Lemma 1. *Consider the problem of convex optimization*

$$f(x) \to \min_{\substack{g(x) \leqslant 0 \\ x \in Q}}.$$

Suppose the Slater condition is satisfied, so there is such a point $\overline{x} \in Q$ that $g(\overline{x}) < 0$, i.e. there exists $\gamma > 0$ such that $g(\overline{x}) = -\gamma < 0$. Then the following estimate holds

$$\lambda^* \leqslant \frac{1}{\gamma}(f(\overline{x}) - \min_{x \in Q} f(x)), \tag{11}$$

where λ^* is a solution of the dual problem $\varphi(\lambda) \to \max\limits_{\lambda \geqslant 0}$.

Proof. Note the following inequality

$$\min_{x\in Q} f(x) = \min_{x\in Q}\left\{ f(x) + \underbrace{\lambda}_{=0} g(x)\right\} \leqslant \max_{\lambda\geqslant 0}\min_{x\in Q}\{f(x)+\lambda g(x)\}$$
$$= \min_{x\in Q}\{f(x)+\lambda^* g(x)\} \leqslant f(\bar{x}) + \lambda^* g(\bar{x}) = f(\bar{x}) + \lambda^*\gamma.$$

Using this inequality one can get

$$\lambda^*\gamma \leqslant f(\overline{x}) - \min_{x\in Q} f(x).$$

□

Thus, by using lemma (1), we can take the upper bound for the dual variable λ as follows:

$$\lambda_{max} = \frac{1}{\gamma}\Big(f(\overline{x}) - \min_{x\in Q} f(x)\Big).$$

3.2 An Estimate of the Accuracy of Solutions for the Proposed Algorithms

To estimate the rate of convergence of the previous Algorithms 1 and 2, we need the following analogue of Theorem 1 from [11].

Theorem 2. *Let $f(x)$ be a μ_f-strongly convex function, the function $g(x)$ satisfies the Lipschitz condition with a constant M_g. Then the function $\varphi(\lambda)$, defined in (7), where $x(\lambda)$ is determined by the condition (8), is an M_g^2/μ_f-smooth function, i.e. the derivative of the function $\varphi(\lambda)$ satisfies the following Lipschitz condition*

$$|\varphi'(\lambda_2) - \varphi'(\lambda_1)| \leqslant L_\varphi\,|\lambda_2-\lambda_1|\,, \tag{12}$$

with a constant $L_\varphi = M_g^2/\mu_f$.

Proof. Let $\lambda_1,\ \lambda_2 \in [\lambda_{min}, \lambda_{max}]$. Define

$$x_1 = \arg\min_{x\in Q}\{f(x)+\lambda_1 g(x)\}\,,\quad x_2 = \arg\min_{x\in Q}\{f(x)+\lambda_2 g(x)\}\,.$$

Since x_1 and x_2 are unique due to the strong convexity of the function f and by using (9), one can get

$$\varphi'(\lambda_1) = g(x_1),\ \varphi'(\lambda_2) = g(x_2).$$

Recall the necessary optimality conditions are

$$\langle \nabla f(x_1) + \lambda_1\nabla g(x_1),\, x_1 - x_2\rangle \leqslant 0,\quad \langle \nabla f(x_2) + \lambda_2\nabla g(x_2),\, x_2 - x_1\rangle \leqslant 0.$$

Summing these inequalities, we get

$$\langle \nabla f(x_1) - \nabla f(x_2),\, x_2 - x_2\rangle \leqslant \langle \lambda_1\nabla g(x_1) - \lambda_2\nabla g(x_2),\, x_2 - x_1\rangle.$$

Due to the strong convexity of $f(x)$, we obtain the following inequality

$$\langle \nabla f(x_2) - \nabla f(x_1),\, x_2 - x_1 \rangle \geqslant \mu_f ||x_2 - x_1||_2^2.$$

Then

$$\begin{aligned} \mu_f ||x_2 - x_1||_2^2 &\leqslant \langle \lambda_1 \nabla g(x_1) - \lambda_2 \nabla g(x_2),\, x_2 - x_1 \rangle \\ &= \underbrace{\lambda_1}_{\geqslant 0} \underbrace{\langle \nabla g(x_1) - \nabla g(x_2),\, x_2 - x_1 \rangle}_{\leqslant 0} + (\lambda_1 - \lambda_2) \langle \nabla g(x_2),\, x_2 - x_1 \rangle \\ &\leqslant |\lambda_1 - \lambda_2| \langle \nabla g(x_2),\, x_2 - x_1 \rangle \leqslant |\lambda_1 - \lambda_2|\, ||\nabla g(x_2)||_2\, ||x_2 - x_1||_2 \\ &\leqslant M_g |\lambda_1 - \lambda_2|\, ||x_2 - x_1||_2, \end{aligned}$$

where $||\nabla g(x_2)||_2 \leqslant M_g$ since g satisfies Lipschitz condition (3).

Thus, for $x_1 \neq x_2$ we get

$$\mu_f ||x_2 - x_1||_2 \leqslant M_g |\lambda_2 - \lambda_1|.$$

As a result, the following estimate holds

$$|\varphi'(\lambda_2) - \varphi'(\lambda_1)| = |g(x_2) - g(x_1)| \leqslant M_g ||x_2 - x_1||_2 \leqslant \frac{M_g^2}{\mu_f} |\lambda_2 - \lambda_1|.$$

□

In order to estimate the accuracy of solutions of the proposed Algorithms 1 and 2, we set the following two lemmas.

Lemma 2. *Suppose the stopping criterion of Algorithm 1 holds for $\lambda = \lambda^N$, then the following inequalities hold*

$$f(x_\delta(\lambda)) - f(x^*) \leqslant \varepsilon + \delta, \quad g(x_\delta(\lambda)) \leqslant \frac{\varepsilon}{\lambda}.$$

For the case $\delta = \varepsilon$ we get

$$f(x_\delta(\lambda)) - f(x^*) \leqslant 2\varepsilon, \quad g(x_\delta(\lambda)) \leqslant \frac{\varepsilon}{\lambda}.$$

Proof. Let λ^* be a solution of the dual problem (7). Denote $x^* = x(\lambda^*)$. Then we get the following relation

$$\begin{aligned} f(x_\delta(\lambda)) + \lambda g(x_\delta(\lambda)) &\leqslant f(x(\lambda)) + \lambda g(x(\lambda)) + \delta = \varphi(\lambda) + \delta \\ &\leqslant \varphi(\lambda^*) + \delta = f(x^*) + \lambda^* \underbrace{g(x^*)}_{\leqslant 0} + \delta \leqslant f(x^*) + \delta. \end{aligned}$$

Consequently,

$$f(x_\delta(\lambda)) - f(x^*) \leqslant -\lambda g(x_\delta(\lambda)) + \delta \leqslant \varepsilon + \delta$$

due to the stopping criterion of Algorithm 1, as required. The inequality $g(x_\delta(\lambda)) \leqslant \frac{\varepsilon}{\lambda}$ follows from the stopping criterion of Algorithm 1 (see item 9).

□

Also an analogue of Lemma 2 takes place.

Lemma 3. *Suppose the stopping criterion of Algorithm 2 holds for* $\lambda = \lambda^N$, *then the following inequalities hold*

$$f(x_\delta(\lambda)) - f(x^*) \leqslant \lambda\varepsilon + \delta, \quad g(x_\delta(\lambda)) \leqslant \varepsilon.$$

For the case $\delta = \varepsilon$ *we get*

$$f(x_\delta(\lambda)) - f(x^*) \leqslant (\lambda + 1)\varepsilon, \quad g(x_\delta(\lambda)) \leqslant \varepsilon.$$

Remark 2. Let us analyze Lemmas 2 and 3. Algorithm 1 (Lemma 2) guarantees the desirable accuracy of the solution with respect to the objective function, but, possibly, unsatisfactory accuracy of the solution with respect to the constraint, as the estimate is huge in case λ is small. Algorithm 2 (Lemma 3) provides the desirable accuracy of the solution with respect to the constraint and, possibly, unsatisfactory accuracy of the solution with respect to the objective function in case λ is huge. So one of the Algorithms 1, 2 surely guarantees the desirable accuracy with respect to both the objective function and the constraint.

3.3 Estimates of the Rate of Convergence for Lipschitz-Continuous Functionals

The idea of the proposed methods is the consistent decrease of the localization interval of the values of the dual variable λ. At each iteration of Algorithms 1 and 2, this interval decreases by 2 times and every time contains λ_*, for which $\lambda_* g(x(\lambda_*)) = 0$ (for Algorithm 1)

$$\lambda_* g(x(\lambda_*)) = \lambda_* \varphi'(\lambda_*) = 0$$

or $g(x(\lambda_*)) = 0$ (for Algorithm 2)

$$g(x(\lambda_*)) = \varphi'(\lambda_*) = 0.$$

By Theorem 2 for all $\lambda_1, \lambda_2 \in [0; \lambda_{\max}]$

$$|\varphi'(\lambda_2) - \varphi'(\lambda_1)| \leqslant \frac{M_g^2}{\mu_f} |\lambda_2 - \lambda_1|, \tag{13}$$

whence

$$|\lambda_2 \varphi'(\lambda_2) - \lambda_1 \varphi'(\lambda_1)| \leqslant \left(|\varphi'(0)| + \frac{M_g^2 \lambda_{\max}}{\mu_f} \right) |\lambda_2 - \lambda_1| = C|\lambda_2 - \lambda_1|, \tag{14}$$

where $C = |\varphi'(0)| + \frac{M_g^2 \lambda_{\max}}{\mu_f}$. Therefore, the achievement of the stopping criterion for Algorithm 2 (item 9) is possible with

$$\lambda_{\max}^N - \lambda_{\min}^N = \frac{\lambda_{\max}}{2^N} \leqslant \frac{\varepsilon}{2C},$$

i.e.

$$N \geqslant \log_2 \frac{2C\lambda_{\max}}{\varepsilon}.$$

So, Algorithm 1 stops after no more than

$$O\left(\log_2 \frac{M_g^2 \lambda_{\max}^2}{\varepsilon \mu_f}\right)$$

iterations. Similarly, if there is $\lambda_* : \varphi'(\lambda_*) = 0$, then (14) means that Algorithm 2 stops after no more than

$$O\left(\log_2 \frac{M_g^2 \lambda_{\max}}{\varepsilon \mu_f}\right)$$

iterations.

Let us analyze the rate of convergence of proposed Algorithms 1 and 2. We need some results from [2] concerning a strongly convex objective function.

Method which guarantees an optimal rate of convergence for the problem (6) is an algorithm based on the restarting of another Adaptive Mirror Descent Algorithm. Information concerning the ordinary Adaptive Mirror Descent Algorithm and the algorithm with its restart can be found in Sect. 4 (Algorithms 3 and 4 respectively). In each iteration of Algorithms 1 and 2 the auxiliary problem

$$x_\delta(\lambda) = \arg\min_{x\in Q}\left\{f(x) + \lambda g(x)\right\}$$

is being solved inexactly with the accuracy δ, which means

$$f(x_\delta(\lambda)) + \lambda g(x_\delta(\lambda)) - f(x^*(\lambda)) + \lambda g(x^*(\lambda)) \leqslant \delta,$$

where the function $f(x) + \lambda g(x)$ is strongly convex and satisfies the Lipschitz condition for any fixed λ due to the properties of the functions $f(x)$ and $g(x)$.

To solve the auxiliary problem of minimization of the functional $F_\lambda(x) = f(x) + \lambda g(x)$, we use the standard gradient method. Let us note an important statement [1]. After k iterations of the standard projected subgradient method the following inequality holds

$$F_\lambda(x^k) - F_\lambda(x^*) \leqslant \frac{2M_{F_\lambda}^2}{k \cdot \mu_f},$$

where $M_{F_\lambda} = \max\{M_f, \lambda \cdot M_g\}$. Due to the strong convexity of f we have

$$F_\lambda(x) \geqslant F_\lambda(x^*) + \langle \nabla F_\lambda(x^*), x - x^*\rangle + \frac{\mu_f}{2}\|x - x^*\|_2^2 \geqslant F_\lambda(x^*) + \frac{\mu_f}{2}\|x - x^*\|_2^2.$$

So,

$$\|x - x^*\|_2^2 \leqslant \frac{2}{\mu_f}\left(F_\lambda(x) - F_\lambda(x^*)\right).$$

Taking $x = x^k$ the following estimate holds

$$\|x - x^*\|_2^2 \leqslant \frac{4M_F^2}{k \cdot \mu_f^2} \leqslant \delta^2.$$

Thus, the required number of iterations does not exceed

$$k = \frac{4M_F^2}{\mu_f^2 \delta^2}.$$

Now by using Theorem 2 and taking into account the complexity $O\left(\log_2(\frac{1}{\varepsilon})\right)$ of the dichotomy in Algorithms 1 and 2, the general complexity is

$$O\left(\frac{1}{\delta^2} log_2 \frac{1}{\varepsilon}\right).$$

Remark 3. If $\delta = \varepsilon$ then the general complexity of Algorithms 1 and 2:

$$O\left(\frac{1}{\varepsilon^2} log_2 \frac{1}{\varepsilon}\right).$$

3.4 Estimate for Composite Formulation

Let us emphasize an important remark. Let f have a Lipschitz-continuous gradient, with a constant L_f

$$||\nabla f(x) - \nabla f(y)||_2 \leqslant L_f ||x - y||_2 \forall x, y \in Q,$$

and g be a so-called simple function, i.e. g is a non-smooth convex function of a simple structure. The latter means that Lebesgue sets

$$\Lambda_y = \{x \in Q : g(x) < y\} \tag{15}$$

have a simple structure. For example, to such problems can be attributed the LASSO problem [3,9,12]:

$$\frac{1}{2}||Ax - b||_2^2 + \lambda ||x||_1 \to \min_{x \in \mathbb{R}^n}, \tag{16}$$

where A is a matrix of $(m \times n)$ dimension, $b \in \mathbb{R}^m$, λ is a regularization parameter and $|| \cdot ||_1$ denotes the standard l_1-norm.

Then we can use the following gradient-type procedure

$$x^{k+1} = \arg\min_{x \in Q} \left\{ \langle \nabla f(x^k), x - x^k \rangle + \lambda g(x) + \frac{L_f}{2} ||x - x^k||_2^2 \right\}. \tag{17}$$

For the method (17) we can achieve $||x - x(\delta)||_2 \leqslant \varepsilon$ after

$$\sqrt{\frac{L_f}{\mu}} \log_2 \frac{1}{\delta}$$

iterations of the method (17) [9]. In such a case, the general complexity of Algorithms 1 and 2:

$$O\left(log_2\frac{1}{\delta}log_2\frac{1}{\varepsilon}\right). \tag{18}$$

The convergence rate is similar in the case when g is a smooth convex function of a simple structure (see (15)). Let g have a Lipschitz-continuous gradient, with a constant L_g

$$||\nabla g(x)-\nabla g(y)||_2 \leqslant L_g||x-y||_2 \forall x,y\in Q$$

and f be a non-smooth convex function. Then we can use the following gradient-type procedure

$$x^{k+1}=\arg\min_{x\in Q}\left\{\langle\lambda\nabla g(x^k),x-x^k\rangle+f(x)+\frac{\lambda L_g}{2}||x-x^k||_2^2\right\}. \tag{19}$$

For the method (19) we can achieve $||x-x(\delta)||_2\leqslant\varepsilon$ after

$$\sqrt{\frac{\lambda L_g}{\mu_f}}\log_2\frac{1}{\delta}$$

iterations of the method (19) and the general complexity (18) for Algorithms 1 and 2.

3.5 The Case of Smooth Functionals

Suppose functions f and g are smooth, i.e. there exist some L_f, L_g such that

$$||\nabla f(x)-\nabla f(y)||_2 \leqslant L_f||x-y||_2 \ \forall x,y\in Q,$$

$$||\nabla g(x)-\nabla g(y)||_2 \leqslant L_g||x-y||_2 \ \forall x,y\in Q.$$

Then the auxiliary problem

$$\arg\min_{x\in Q}F_\lambda(x),$$

where $F_\lambda(x)=f(x)+\lambda g(x)$, is also smooth and it can be solved, for example, with Gradient Descent [9]

$$x^{k+1}=x^k-\alpha\nabla F_\lambda(x^k).$$

Note that F_λ is a μ_f-strongly convex function.

In such a case, the following estimate for the rate of convergence holds ([6], [9])

$$||x^k-x(\delta)||_2^2\leqslant||x^0-x(\delta)||_2^2\left(1-\frac{\mu_f}{\max\{L_f,\lambda L_g\}}\right)^k.$$

It means that the complexity of Algorithms 1 and 2 is (18). For $\delta=\varepsilon$ the estimate (18) is

$$O\left(log_2^2\frac{1}{\varepsilon}\right).$$

4 Comparison with Mirror Descent Algorithms

In this section, we compare the proposed methods with two variants of the Mirror Descent Algorithm. These are the classical variant and the one based on the restart method. Let us, according to [2], present basic information concerning Mirror Descent Algorithms. Assume that there exists a constant $\Theta_0 > 0$, that $\frac{1}{2}\|x - x^*\|_2^2 \leq \Theta_0^2$. If there is a set of solutions of the problem $\{x_i^*\}$, assume that

$$\min_{x^* \in \{x_i^*\}} \frac{1}{2}\|x - x^*\|_2^2 \leq \Theta_0^2.$$

The standard definition of the mirror descent operator with Euclidean proximal setup is defined as

$$Mirr_x(p) = \arg\min_{v \in Q} \left\{ \langle p, v \rangle + \frac{1}{2}\|x - v\|_2^2 \right\} \text{ for each } x \in Q \text{ and } p \in E^*,$$

and assume that it is easily computable.

Algorithm 3. Adaptive Mirror Descent Algorithm.

```
Require: $\varepsilon > 0, \Theta_0$ s.t. $\frac{1}{2}\|x - x^*\|_2^2 \leqslant \Theta_0^2$.
$x^0 = argmin_{x \in Q} \frac{1}{2}\|x - x^*\|_2^2$
$I =: \emptyset$
$N \leftarrow 0$
repeat
    if $g(x^N) \leqslant \varepsilon$ then
        $M_N = ||\nabla f(x^N)||_2$, $h_N = \frac{\varepsilon}{M_N^2}$
        $x^{N+1} = Mirr_{x^N}(h_N \nabla f(x^N))$ "productive step"
        $N \rightarrow I$
    else
        $M_N = ||\nabla g(x^N)||_2$, $h_N = \frac{\varepsilon}{M_N^2}$
        $x^{N+1} = Mirr_{x^N}(h_N \nabla g(x^N))$ "non-productive step"
    end if
    $N \leftarrow N + 1$
until $\sum_{j=0}^{N-1} \frac{1}{M_j^2} \geqslant 2\frac{\Theta_0^2}{\varepsilon^2}$
Ensure: $\bar{x}^N := \frac{\sum_{k \in I} x^k h_k}{\sum_{k \in I} h_k}$
```

Theorem 3. *Let the functionals f and g satisfy the Lipschitz condition with constants M_f and M_g respectively. Then Algorithm 3 works no more than*

$$N = \left\lceil \frac{2\max\{M_f^2, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil$$

iterations, and the point $\overline{x}^N$ *is a* ε*-solution of* (6)*. It means that*

$$f(\overline{x}^k) - f(x^*) \leq \varepsilon, \quad g(\overline{x}^k) \leq \varepsilon. \tag{20}$$

Consider the case of μ-strong convex f and g. We need to modify some proposed assumptions. Assume that

$$x_0 = \arg\min_{x \in Q} \frac{1}{2}\|x - x^*\|_2^2, \quad \frac{1}{2}\|x - x^*\|_2^2 \leq \frac{\Omega}{2} \quad \forall x \in Q : \|x\|_2 \leq 1,$$

where Ω is some known constant. Suppose that there exists some initial starting point $x_0 \in Q$ and a number $R_0 > 0$ such that $\|x_0 - x^*\|_2^2 \leq R_0^2$.

Algorithm 4. Adaptive Mirror Descent Algorithm for Strongly Convex Functions (with restart of Algorithm 3).

```
Require: accuracy $\varepsilon > 0$; starting point $x_0$; $\Omega$ s.t. $\frac{1}{2}\|x - x^*\|_2^2 \leq \frac{\Omega}{2}$ $\forall x \in Q : \|x\|_2 \leq 1$;
    strong convexity parameter $\mu$; $R_0$ s.t. $\|x_0 - x^*\|_2^2 \leq R_0^2$.
Set $d_0(x) = \frac{1}{2}\|\left(\frac{x - x_0}{R_0}\right) - x^*\|_2^2$.
Set $p = 1$.
repeat
    Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
    Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
    Set $x_p$ as the output of Algorithm 3 with accuracy $\varepsilon_p$, prox-function $d_{p-1}(\cdot)$ and
    $\frac{\Omega}{2}$ as $\Theta_0^2$.
    $d_p(x) \leftarrow \frac{1}{2}\|\left(\frac{x - x_p}{R_p}\right) - x^*\|_2^2$.
    Set $p = p + 1$.
until $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.
Ensure: $x^p$.
```

Theorem 4. *Assume that* f *and* g *satisfy the Lipschitz condition with constants* M_f *and* M_g *respectively. Then solving the* μ*-strongly convex problem* (6)*, Algorithm 4 works no more than*

$$k = \left\lceil log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil + \frac{32\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon}$$

iterations. The output point x_p *of Algorithm 4 is satisfied to* (20) *and the following inequality holds*

$$\|x_p - x^*\|_2^2 \leq \frac{2\varepsilon}{\mu}.$$

5 Numerical Experiments

To compare Algorithms 1, 2 and 4, a series of numerical experiments were carried out. Consider three different examples of strongly convex, Lipschitz-continuous objective functions, as follows

Example 1.

$$f(x) = x_1^2 + \sum_{i=1}^{n} i x_i^2 + \frac{1}{100} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j \right)^2 .$$

Example 2.

$$f(x) = \sum_{i=1}^{n-1} i x_i^2 + \sum_{i=1}^{n-2} \left(x_i + x_{i+1} + x_{i+2} \right)^2 .$$

Example 3.

$$f(x) = \sum_{i=1}^{n} i x_i^4 + \frac{1}{2} \|x\|_2^2 .$$

The functional constraint has the next form: $g(x) = \max\limits_{1 \leq i \leq m} \{g_i(x)\}$, where

$$g_i((x_1, \ldots, x_n)) = \langle a_i x, x \rangle - 5,$$

a_i^T $(i = 1, \ldots, m)$ are the rows in the matrix $A \in \mathbb{R}^{m \times n}$ with entries drawn from the discrete uniform distribution in the half open interval $[1, 6)$.

Let us choose the set $Q = \{x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n \,;\, x_1^2 + x_2^2 + ... + x_n^2 \leq 1\}$. For Algorithms 1 and 2, we choose $\lambda_{min} = 0, \lambda_{max} = \frac{f(\bar{x})}{-g(\bar{x})}$, where $\bar{x}$ is an arbitrary point such that $g(\bar{x}) < 0$. For Algorithm 4 we choose standard Euclidean proximal setup as prox-function, starting point $x_0 = \frac{(1, \ldots, 1)}{\sqrt{n}}$, $\Theta_0 = \sqrt{2}$ (i.e. $\Omega = 4$) and $R_0 = 1$.

For $\varepsilon = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$ the results of the work of Algorithms 1, 2 and 4, for Examples 1 and 2, when $n = 200, m = 100$, are represented in Figs. 1 and 2 below. For Example 3, when $n = 1000$ and $m = 100$, they are represented in Fig. 3. These results demonstrate the comparison of the running time (in seconds) for each algorithm, with different accuracy ε.

All experiments were implemented in Python 3.4, on a computer fitted with Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s). The RAM of the computer is 8 GB.

In general, from all experiments conducted, we can see that Algorithm 1 is the best algorithm, the efficiency of this algorithm is represented by its very high execution speed, where by this algorithm one needs a few seconds to achieve the solution and to reach its stopping criterion. In some details, from Fig. 1 and Fig. 2, for Examples 1 and 2 when $n = 200, m = 100$, one can see that, according to the running time of each algorithm, Algorithm 1 works better than Algorithm 2, which works better than Algorithm 4. We note that the running time of Algorithm 4 is very long compared with the running time of Algorithms 1 and 2. Therefore, for the objective functions in Examples 1 and 2 (quadratic functions), we can see that Algorithm 4 works badly, unlike Algorithm 1. For Example 3 when $n = 1000, m = 100$, from Fig. 3, we can see that Algorithm 1 is still the best, but now Algorithm 4 works better than Algorithm 2. We note that the difference between the running time of Algorithms 1 and 4 is very small, but it is very long compared with the running time of Algorithm 2.

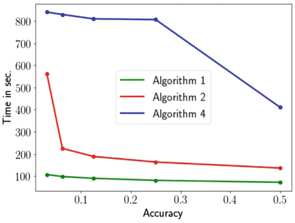

Fig. 1. Example 1, $n = 200$.

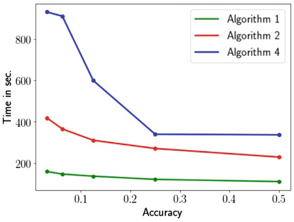

Fig. 2. Example 2, $n = 200$.

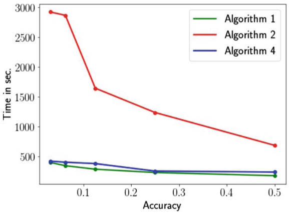

Fig. 3. Example 3, $n = 1000$.

References

1. Aravkin A.Y., Burke J.V., Drusvyatskiy D.: Convex Analysis and Nonsmooth Optimization (2017). https://sites.math.washington.edu/~burke/crs/516/notes/graduate-nco.pdf
2. Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F., Titov, A.: Mirror descent and convex optimization problems with non-smooth inequality constraints. In: Giselsson, P., Rantzer, A. (eds.) Large-Scale and Distributed Optimization. LNM, vol. 2227, pp. 181–213. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97478-1_8
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
4. Ben-Tal, A., Nemirovski, A.: Robust truss topology design via semidefinite programming. SIAM J. Optim. **7**(4), 991–1016 (1997)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2004)
6. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015). https://arxiv.org/pdf/1405.4980.pd
7. Danskin, J.M.: The theory of Max-Min, with applications. J. SIAM Appl. Math. **14**(4) (1966)
8. Demyanov, V.F., Malozemov, V.N.: Introduction to Minimax. Nauka, Moscow (1972). (in Russian)
9. Gasnikov, A.V.: Modern numerical optimization methods. The method of universal gradient descent (2018). (in Russian). https://arxiv.org/ftp/arxiv/papers/1711/1711.00394.pdf
10. Ivanova, A., Gasnikov, A., Nurminski, E., Vorontsova, E.: Walrasian equilibrium and centralized distributed optimization from the point of view of modern convex optimization methods on the example of resource allocation problem (2019). (in Russian). https://arxiv.org/pdf/1806.09071.pdf
11. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
12. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
13. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Massachusetts (2004)
14. Shpirko, S., Nesterov, Y.: Primal-dual subgradient methods for huge-scale linear conic problem. SIAM J. Optim. **24**(3), 1444–1457 (2014)
15. Vasilyev, F.: Optimization Methods. Fizmatlit, Moscow (2002). (in Russian)