Analogues of Switching Subgradient Schemes for Relatively Lipschitz-Continuous Convex Programming Problems

Alexander A. Titov[1,3]($^{(\boxtimes)}$) , Fedor S. Stonyakin[1,2] ,
Mohammad S. Alkousa[1,3] , Seydamet S. Ablaev[2] ,
and Alexander V. Gasnikov[1]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
{a.a.titov,mohammad.alkousa}@phystech.edu, gasnikov@yandex.ru
[2] V. I. Vernadsky Crimean Federal University, Simferopol, Russia
fedyor@mail.ru, seydamet.ablaev@yandex.ru
[3] National Research University, Higher School of Economics, Moscow, Russia

Abstract. Recently some specific classes of non-smooth and non-Lipsch-itz convex optimization problems were considered by Yu. Nesterov and H. Lu. We consider convex programming problems with similar smoothness conditions for the objective function and functional constraints. We introduce a new concept of an inexact model and propose some analogues of switching subgradient schemes for convex programming problems for the relatively Lipschitz-continuous objective function and functional constraints. Some class of online convex optimization problems is considered. The proposed methods are optimal in the class of optimization problems with relatively Lipschitz-continuous objective and functional constraints.

Keywords: Convex programming problem · Switching subgradient scheme · Relative lipschitz-continuity · Inexact model · Stochastic mirror descent · Online optimization problem

1   Introduction

Different relaxations of the classical smoothness conditions for functions are interesting for a large number of modern applied optimization problems. In particular, in [2] there were proposed conditions of relative smoothness of the objective function, which mean the replacement of the classic Lipschitz condition by the following weaker version

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + LV_d(y, x), \tag{1}$$

to hold for any x, y from the domain of the objective function f and some $L > 0$; $V_d(y, x)$ represents an analogue of the distance between the points x and y (often called the *Bregman divergence*). Such a distance is widely used in various fields of science, in particular in mathematical optimization. Usually, the *Bregman divergence* is defined on the base of the auxiliary 1-strongly convex and continuously-differentiable function $d : Q \subset \mathbb{R}^n \to \mathbb{R}$ (*distance generating function*) as follows

$$V_d(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle \quad \forall x, y \in Q, \tag{2}$$

where Q is a convex closed set, $\langle \cdot, \cdot \rangle$ is a scalar product in $\mathbb{R}^n$. In particular, for the Euclidean setting of the problem, we have $d(x) = \frac{1}{2}\|x\|_2^2$ and $V_d(y, x) = d(y - x) = \frac{1}{2}\|y - x\|_2^2$ for arbitrary $x, y \in Q$. However, in many applications, it often becomes necessary to use non-Euclidean norms. Moreover, the considered condition of relative smoothness in [2,16] implies only the convexity (but not strong convexity) of the distance generating function d. As shown in [16], the concept of relative smoothness makes it possible to apply a variant of the gradient method to some problems which were previously being solved only by interior-point methods. In particular, we talk about the well-known problem of construction of an optimal ellipsoid which covers a given set of points. This problem is important in the field of statistics and data analysis.

A similar approach to the Lipschitz property and non-smooth problems was proposed in [17] (see also [24]). This approach is based on an analogue of the Lipschitz condition for the objective function $f : Q \to \mathbb{R}$ with Lipschitz constant $M_f > 0$, which involves replacing the boundedness of the norm of the subgradient, i.e. $\|\nabla f(x)\|_* \leq M_f$, with the so-called *relative Lipschitz condition*

$$\|\nabla f(x)\|_* \leq \frac{M_f \sqrt{2V_d(y, x)}}{\|y - x\|} \quad \forall x, y \in Q, \ y \neq x,$$

where $\|\cdot\|_*$ denotes the conjugate norm, see Sect. 2 below. Moreover, the distance generating function d must not necessarily be strongly convex. In [17] there were proposed deterministic and stochastic Mirror Descent algorithms for optimization problems with convex relatively Lipschitz-continuous objective functionals. Note that some applications of relative Lipschitz-continuity to the well-known classical support vector machine (SVM) problem and to the problem of minimizing the maximum of convex quadratic functions (intersection of m ellipsoids problem in $\mathbb{R}^n$) were discussed in [17].

In this paper we propose a new concept of an inexact model for objective functional and functional constraint. More precisely, we introduce some analogues of the concepts of an inexact oracle [8] and an inexact model [32] for objective functionals. However, unlike [8,32], we do not generalize the smoothness condition. We relax the Lipschitz condition and consider a recently proposed generalization of *relative Lipschitz-continuity* [17,24]. We propose some optimal Mirror Descent methods, in different settings of Relatively Lipschitz-continuous convex optimization problems.

The Mirror Descent method originated in the works of A. Nemirovski and D. Yudin more than 30 years ago [21,22] and was later analyzed in [5]. It can be considered as the non-Euclidean extension of subgradient methods. The method was used in many applications [19,20,31]. Standard subgradient methods employ the Euclidean distance function with a suitable step-size in the projection step. The Mirror Descent extends the standard projected subgradient methods by employing a nonlinear distance function with an optimal step-size in the nonlinear projection step [18]. The Mirror Descent method not only generalizes the standard subgradient descent method, but also achieves a better convergence rate and it is applicable to optimization problems in Banach spaces, while the subgradient descent is not [9]. Also, in some works [4,10,22] there was proposed an extension of the Mirror Descent method for constrained problems.

Also, in recent years, online convex optimization (OCO) has become a leading online learning framework, due to its powerful modeling capability for a lot of problems from diverse domains. OCO plays a key role in solving problems where statistical information is being updated [13,14]. There are many examples of such problems: Internet networks, consumer data sets or financial markets, machine learning applications, such as adaptive routing in networks, dictionary learning, classification and regression (see [33] and references therein). In recent years, methods for solving online optimization problems have been actively developed, in both deterministic and stochastic settings [7,12,15,25]. Among them one can mention the Mirror Descent method for the deterministic setting of the problem [26,30] and for the stochastic setting [1,11,34,35], which allows to solve problems for an arbitrary distance function.

This paper is devoted to Mirror Descent methods for convex programming problems with a relatively Lipschitz-continuous objective function and functional constraints. It consists of an introduction and 5 main sections. In Sect. 2 we consider the problem statement and define the concept of an inexact (δ, ϕ, V)–model for the objective function. Also, we propose some modifications of the Mirror Descent method for the concept of Model Generality. Section 3 is devoted to some special cases of problems with the properties of relative Lipschitz continuity, here we propose two versions of the Mirror Descent method in order to solve the problems under consideration. In Sects. 4 and 5 we consider the stochastic and online (OCO) setting of the optimization problem respectively. In Sect. 6 one can find numerical experiments which demonstrate the efficiency of the proposed methods.

The contribution of the paper can be summarized as follows:

- Continuing the development of Yurii Nesterov's ideas in the direction of the relative smoothness and non-smoothness [24], we introduced the concept of an inexact (δ, ϕ, V)–model of the objective function. For the proposed model we proposed some variants of the well-known Mirror Descent method, which provides an $(\varepsilon + \delta)$–solution of the optimization problem, where ε is the controlled accuracy. There was considered the applicability of the proposed method to the case of the stochastic setting of the considered optimization problem.
- We also considered a special case of the relative Lipschitz condition for the objective function. The proposed Mirror Descent algorithm was specified for the case of such functions. Furthermore, there was introduced one more modification of the algorithm with another approach to the step selection. There was also considered the possibility of applying the proposed methods to the case of several functional constraints.
- We considered an online optimization problem and proposed a modification of the Mirror Descent algorithm for such a case. Moreover, there were conducted some numerical experiments which demonstrate the effectiveness of the proposed methods.

2   Inexact Model for Relative Non-smooth Functionals and Mirror Descent Algorithm

Let $(E, \|\cdot\|)$ be a normed finite-dimensional vector space and E^* be the conjugate space of E with the norm:

$$\|y\|_* = \max_x \{\langle y, \, x \rangle, \|x\| \leq 1\},$$

where $\langle y, \, x \rangle$ is the value of the continuous linear functional y at $x \in E$.

Let $Q \subset E$ be a (simple) closed convex set. Consider two subdifferentiable functions $f, g : Q \to \mathbb{R}$. In this paper we consider the following optimization problem

$$f(x) \to \min_{x \in Q, \, g(x) \leq 0}. \tag{3}$$

Let $d : Q \to \mathbb{R}$ be any convex (not necessarily strongly-convex) differentiable function, we will call it *reference function*. Suppose we have a constant $\Theta_0 > 0$, such that $d(x^*) \leq \Theta_0^2$, where x^* is a solution of (3). Note that if there is a set, $X_* \subset Q$, of optimal points for the problem (3), we may assume that

$$\min_{x^* \in X_*} d(x^*) \leq \Theta_0^2.$$

Let us introduce some generalization of the concept of relative Lipschitz continuity [24]. Consider one more auxiliary function $\phi : \mathbb{R} \to \mathbb{R}$, which is strictly increasing and satisfies $\phi(0) = 0$. Clearly, due to the strict monotonicity of $\phi(\cdot)$, there exists the inverse function $\phi^{-1}(\cdot)$.

Definition 1. *Let $\delta > 0$. We say that f and g admit the (δ, ϕ, V)–model at the point $y \in Q$ if*

$$f(x) + \psi_f(y, x) \le f(y), \quad -\psi_f(y, x) \le \phi_f^{-1}(V_d(y, x)) + \delta \tag{4}$$

$$g(x) + \psi_g(y, x) \le g(y), \quad -\psi_g(y, x) \le \phi_g^{-1}(V_d(y, x)) + \delta, \tag{5}$$

where $\psi_f(\cdot, x)$ and $\psi_g(\cdot, x)$ are convex functions for fixed x and $\psi_f(x, x) = \psi_g(x, x) = 0$ for all $x \in Q$.

Let $h > 0$. For problems with a (δ, ϕ, V)–model, the proximal mapping operator (Mirror Descent step) is defined as follows

$$Mirr_h(x, \psi) = \arg\min_{y \in Q} \left\{ \psi(y, x) + \frac{1}{h} V_d(y, x) \right\}.$$

The following lemma describes the main property of this operator.

Lemma 1 (Main Lemma). *Let f be a convex function, which satisfies (4), $h > 0$ and $\tilde{x} = hMirr_h(x, \psi_f)$. Then for any $y \in Q$*

$$h(f(x) - f(y)) \le -h\psi_f(y, x) \le \phi_f^*(h) + V_d(y, x) - V_d(y, \tilde{x}) + h\delta,$$

where ϕ_f^* is the conjugate function of ϕ_f.

Proof. From the definition of $\tilde{x}$

$$\tilde{x} = hMirr_h(x, \psi_f) = \arg\min_{y \in Q} \left\{ h\psi_f(y, x) + V_d(y, x) \right\},$$

for any $y \in Q$, we have $h\psi_f(y, x) - h\psi_f(\tilde{x}, x) + \langle \nabla d(\tilde{x}) - \nabla d(x), y - \tilde{x} \rangle \ge 0$.
     Further, $h(f(x) - f(y)) \le -h\psi_f(y, x) \le$

$$\le -h\psi_f(\tilde{x}, x) + \langle \nabla d(\tilde{x}) - \nabla d(x), y - \tilde{x} \rangle$$
$$= -h\psi_f(\tilde{x}, x) + V_d(y, x) - V_d(y, \tilde{x}) - V_d(\tilde{x}, x) + h\delta$$
$$\le h\phi_f^{-1}(V_d(\tilde{x}, x)) + V_d(y, x) - V_d(y, \tilde{x}) - V_d(\tilde{x}, x) + h\delta$$
$$\le \phi_f^*(h) + \phi_f(\phi_f^{-1}(V_d(\tilde{x}, x))) + V_d(y, x) - V_d(y, \tilde{x}) - V_d(\tilde{x}, x) + h\delta$$
$$= \phi_f^*(h) + V_d(\tilde{x}, x) + V_d(y, x) - V_d(y, \tilde{x}) - V_d(\tilde{x}, x) + h\delta$$
$$= \phi_f^*(h) + V_d(y, x) - V_d(y, \tilde{x}) + h\delta.$$

For problem (3) with an inexact (δ, ϕ, V)–model, we consider a Mirror Descent algorithm, listed as Algorithm 1 below. For this proposed algorithm, we will call step k productive if $g(x^k) \le \varepsilon$, and non-productive if the reverse inequality $g(x^k) > \varepsilon$ holds. Let I and $|I|$ denote the set of indexes of productive steps and their number, respectively. Similarly, we use the notation J and $|J|$ for non-productive steps.
     Let x^* denote the exact solution of the problem (3). The next theorem provides the complexity and quality of the proposed Algorithm 1.

Theorem 1 (Modified MDA for Model Generality). *Let f and g be convex functionals, which satisfy (4), (5) respectively and $\varepsilon > 0, \delta > 0$ be fixed positive numbers. Assume that $\Theta_0 > 0$ is a known constant such that $d(x^*) \leq \Theta_0^2$. Then, after the stopping of Algorithm 1, the following inequalities hold:*

$$f(\widehat{x}) - f(x^*) \leq \varepsilon + \delta \quad and \quad g(\widehat{x}) \leq \varepsilon + \delta.$$

Algorithm 1. Modified MDA for (δ, ϕ, V)–model.

Require: $\varepsilon > 0, \delta > 0, h^f > 0, h^g > 0, \Theta_0 : d(x^*) \leq \Theta_0^2$.
1: $x^0 = \arg\min_{x \in Q} d(x)$.
2: $I =: \emptyset$ and $J =: \emptyset$
3: $N \leftarrow 0$
4: **repeat**
5:     **if** $g\left(x^N\right) \leq \varepsilon + \delta$ **then**
6:         $x^{N+1} = Mirr_{h^f}\left(x^N, \psi_f\right),$ "productive step"
7:         $N \rightarrow I$
8:     **else**
9:         $x^{N+1} = Mirr_{h^g}\left(x^N, \psi_g\right),$ "non-productive step"
10:        $N \rightarrow J$
11:    **end if**
12:    $N \leftarrow N + 1$
13: **until** $\Theta_0^2 \leq \varepsilon \left(|J|h^g + |I|h^f\right) - |J|\phi_g^*(h^g) - |I|\phi_f^*(h^f)$.
Ensure: $\widehat{x} := \frac{1}{|I|} \sum\limits_{k \in I} x^k$.

Proof. By Lemma 1, we have for all $k \in I$ and $y \in Q$

$$h^f\left(f(x^k) - f(y)\right) \leq \phi_f^*(h^f) + V_d(y, x^k) - V_d(y, x^{k+1}) + h^f\delta. \quad (6)$$

Similarly, for all $k \in J$ and $y \in Q$

$$h^g\left(g(x^k) - g(y)\right) \leq \phi_g^*(h^g) + V_d(y, x^k) - V_d(y, x^{k+1}) + h^g\delta. \quad (7)$$

Summing up these inequalities over productive and non-productive steps, we get

$$\sum_{k \in I} h^f\left(f(x^k) - f(x^*)\right) + \sum_{k \in J} h^g\left(g(x^k) - g(x^*)\right)$$

$$\leq \sum_{k \in I} \phi_f^*(h^f) + \sum_{k \in J} \phi_g^*(h^g) + \sum_{k}\left(V_d(x^*, x^k) - V_d(x^*, x^{k+1})\right) + \sum_{k \in I} h^f\delta + \sum_{k \in J} h^g\delta$$

$$\leq \sum_{k \in I} \phi_f^*(h^f) + \sum_{k \in J} \phi_g^*(h^g) + \Theta_0^2 + \sum_{k \in I} h^f\delta + \sum_{k \in J} h^g\delta.$$

Since for any $k \in J$, $g(x^k) - g(x^*) > \varepsilon + \delta$, we have

$$\sum_{k \in I} h^f \left(f(x^k) - f(x^*) \right) \leq \sum_{k \in I} \phi_f^*(h^f) + \sum_{k \in J} \phi_g^*(h^g) + \Theta_0^2 - \varepsilon \sum_{k \in J} h^g + \sum_{k \in I} h^f \delta$$

$$= |I| \left(\phi_f^*(h^f) + \delta h^f \right) + |J| \phi_g^*(h^g) - |J| h^g \varepsilon + \Theta_0^2 \leq \varepsilon |I| h^f + \delta |I| h^f.$$

So, for $\widehat{x} := \frac{1}{|I|} \sum_{k \in I} x^k$, after the stopping criterion of Algorithm 1 is satisfied, the following inequalities hold

$$f(\widehat{x}) - f(x^*) \leq \varepsilon + \delta \quad \text{and} \quad g(\widehat{x}) \leq \varepsilon + \delta.$$

3   The Case of Relatively Lipschitz-Continuous Functionals

Suppose hereinafter that the objective function f and the constraint g satisfy the so-called relative Lipschitz condition, with constants $M_f > 0$ and $M_g > 0$, i.e. the functions ϕ_f^{-1} and ϕ_g^{-1} from (4) and (5) are modified as follows:

$$\phi_f^{-1} \left(V_d(y, x) \right) = M_f \sqrt{2V_d(y, x)}, \tag{8}$$

$$\phi_g^{-1} \left(V_d(y, x) \right) = M_g \sqrt{2V_d(y, x)}. \tag{9}$$

Note that the functions f, g must still satisfy the left inequalities in (4), (5):

$$f(x) + \psi_f(y, x) \leq f(y), \quad -\psi_f(y, x) \leq M_f \sqrt{2V_d(y, x)} + \delta; \tag{10}$$

$$g(x) + \psi_g(y, x) \leq g(y), \quad -\psi_g(y, x) \leq M_g \sqrt{2V_d(y, x)} + \delta, \tag{11}$$

For this particular case we say that f and g admit the (δ, M_f, V)– and (δ, M_g, V)–model at each point $x \in Q$ respectively. The following remark provides an explicit form of ϕ_f, ϕ_g and their conjugate functions ϕ_f^*, ϕ_g^*.

Remark 1. Let $M_f > 0$ and $M_g > 0$. Then functions ϕ_f and ϕ_g which correspond to (8) and (9) are defined as follows:

$$\phi_f(t) = \frac{t^2}{2M_f^2}, \quad \phi_g(t) = \frac{t^2}{2M_g^2}.$$

Their conjugate functions have the following form:

$$\phi_f^*(y) = \frac{y^2 M_f^2}{2}, \tag{12}$$

$$\phi_g^*(y) = \frac{y^2 M_g^2}{2}. \tag{13}$$

For the case of a relatively Lipschitz-continuous objective function and constraint, we consider a modification of Algorithm 1, the modified algorithm is listed as Algorithm 2, below. The difference between Algorithms 1 and 2 is represented in the control of productivity and the stopping criterion.

For the proposed Algorithm 2, we have the following theorem, which provides an estimate of its complexity and the quality of the solution of the problem.

Theorem 2. *Let f and g be convex functions, which satisfy* (10) *and* (11) *for $M_f > 0$ and $M_g > 0$. Let $\varepsilon > 0, \delta > 0$ be fixed positive numbers. Assume that $\Theta_0 > 0$ is a known constant such that $d(x^*) \leq \Theta_0^2$. Then, after the stopping of Algorithm 2, the following inequalities hold:*

$$f(\widehat{x}) - f(x^*) \leq M_f \varepsilon + \delta \quad and \quad g(\widehat{x}) \leq M_g \varepsilon + \delta.$$

Algorithm 2. Mirror Descent for Relatively Lipschitz-continuous functions, version 1.

Require: $\varepsilon > 0, \delta > 0, M_f > 0, M_g > 0, \Theta_0 : d(x^*) \leq \Theta_0^2$
1: $x^0 = \arg\min_{x \in Q} d(x)$.
2: $I =: \emptyset$
3: $N \leftarrow 0$
4: **repeat**
5:     **if** $g\left(x^N\right) \leq M_g \varepsilon + \delta$ **then**
6:         $h^f = \frac{\varepsilon}{M_f}$,
7:         $x^{N+1} = Mirr_{h^f}\left(x^N, \psi_f\right),$ "productive step"
8:         $N \to I$
9:     **else**
10:       $h^g = \frac{\varepsilon}{M_g}$,
11:       $x^{N+1} = Mirr_{h^g}\left(x^N, \psi_g\right),$ "non-productive step"
12:     **end if**
13:     $N \leftarrow N + 1$
14: **until** $N \geq \frac{2\Theta_0^2}{\varepsilon^2}$.
Ensure: $\widehat{x} := \frac{1}{|I|} \sum_{k \in I} x^k$.

Proof. By Lemma 1, we have

$$\sum_{k \in I} h^f \left(f(x^k) - f(x^*)\right) + \sum_{k \in J} h^g \left(g(x^k) - g(x^*)\right) \leq \sum_{k \in I} \phi_f^*(h^f) + \sum_{k \in J} \phi_g^*(h^g)$$
$$+ \Theta_0^2 + \sum_{k \in I} h^f \delta + \sum_{k \in J} h^g \delta$$

Since for any $k \in J$, $g(x^k) - g(x^*) > M_g \varepsilon + \delta$, we have

$$\sum_{k \in I} h^f \left(f(x^k) - f(x^*)\right) \leq \sum_{k \in I} \phi_f^*(h^f) + \sum_{k \in J} \phi_g^*(h^g) + \Theta_0^2 - M_g \varepsilon \sum_{k \in J} h^g + \sum_{k \in I} h^f \delta$$
$$= |I|(\phi_f^*(h^f) + \delta h^f) + |J|\phi_g^*(h^g) - |J|\varepsilon^2 + \Theta_0^2.$$

Taking into account the explicit form of the conjugate functions (12), (13) one can get:

$$\sum_{k \in I} h^f \left(f(x^k) - f(x^*) \right) \leq |I| \left(\frac{M_f^2(h^f)^2}{2} + \delta h^f \right) + |J| \frac{M_g^2(h^g)^2}{2} - |J|\varepsilon^2 + \Theta_0^2$$

$$= |I| \left(\frac{\varepsilon^2}{2} + \delta h^f \right) + |J| \frac{\varepsilon^2}{2} - |J|\varepsilon^2 + \Theta_0^2$$

$$\leq M_f \varepsilon |I| h^f + \delta |I| h^f,$$

supposing that the stopping criterion is satisfied.

So, for the output value of the form $\widehat{x} = \frac{1}{|I|} \sum_{k \in I} x^k$, the following inequalities hold:

$$f(\widehat{x}) - f(x^*) \leq M_f \varepsilon + \delta \quad \text{and} \quad g(\widehat{x}) \leq M_g \varepsilon + \delta.$$

Also, for the case of a relatively Lipschitz-continuous objective function and constraint, we consider another modification of Algorithm 1, which is listed as the following Algorithm 3. Note that the difference between Algorithm 2 and Algorithm 3 lies in the choice of steps h^f, h^g and the stopping criterion.

Algorithm 3. Mirror Descent for Relatively Lipschitz-continuous functions, version 2.

Require: $\varepsilon > 0, \delta > 0, M_f > 0, M_g > 0, \Theta_0 : d(x^*) \leq \Theta_0^2.$
1: $x^0 = \arg\min_{x \in Q} d(x).$
2: $I =: \emptyset$ and $J =: \emptyset$
3: $N \leftarrow 0$
4: **repeat**
5:     **if** $g\left(x^N\right) \leq \varepsilon + \delta$ **then**
6:         $h^f = \frac{\varepsilon}{M_f^2},$
7:         $x^{k+1} = Mirr_{h^f}\left(x^N, \psi_f\right),$ "productive step"
8:         $N \rightarrow I$
9:     **else**
10:         $h^g = \frac{\varepsilon}{M_g^2},$
11:         $x^{N+1} = Mirr_{h^g}\left(x^N, \psi_g\right),$ "non-productive step"
12:         $N \rightarrow J$
13:     **end if**
14:     $N \leftarrow N + 1$
15: **until** $\frac{2\Theta_0^2}{\varepsilon^2} \leq \frac{|I|}{M_f^2} + \frac{|J|}{M_g^2}.$
Ensure: $\widehat{x} := \frac{1}{|I|} \sum_{k \in I} x^k.$

By analogy with the proof of Theorem 2 one can obtain the following result concerning the quality of the convergence of the proposed Algorithm 3.

Theorem 3. *Let f and g be convex functions, which satisfy* (10) *and* (11) *for $M_f > 0$ and $M_g > 0$. Let $\varepsilon > 0, \delta > 0$ be fixed positive numbers. Assume that $\Theta_0 > 0$ is a known constant such that $d(x^*) \leq \Theta_0^2$. Then, after the stopping of Algorithm 3, the following inequalities hold:*

$$f(\widehat{x}) - f(x^*) \leq \varepsilon + \delta \quad and \quad g(\widehat{x}) \leq \varepsilon + \delta.$$

Moreover, the required number of iterations of Algorithm 3 does not exceed

$$N = \frac{2M^2 \Theta_0^2}{\varepsilon^2}, \quad where \ M = \max\{M_f, M_g\}.$$

Remark 2. Clearly, Algorithms 2 and 3 are optimal in terms of the lower bounds [22]. More precisely, let us understand hereinafter the optimality of the Mirror Descent methods as the complexity $O(\frac{1}{\varepsilon^2})$ (it is well-known that this estimate is optimal for Lipschitz-continuous functionals [22]).

Remark 3 (The case of several functional constraints). Let us consider a set of convex functions f and $g_p : Q \to \mathbb{R}$, $p \in [m] \stackrel{\text{def}}{=} \{1, 2, \ldots, m\}$. We will focus on the following constrained optimization problem

$$\min\{f(x): \ x \in Q \ \text{and} \ g_p(x) \leq 0 \ \text{for all} \ p \in [m]\}. \tag{14}$$

It is clear that instead of a set of functionals $\{g_p(\cdot)\}_{p=1}^m$ we can consider one functional constraint $g : Q \to \mathbb{R}$, such that $g(x) = \max_{p \in [m]}\{g_p(x)\}$. Therefore, by this setting, problem (14) will be equivalent to the problem (3).

Assume that for any $p \in [m]$, the functional g_p satisfies the following condition

$$-\psi_{g_p}(y, x) \leq M_{g_p}\sqrt{2V_d(y, x)} + \delta.$$

For problem (14), we propose a modification of Algorithms 2 and 3 (the modified algorithms are listed as Algorithm 6 and 7 in [29], Appendix A). The idea of the proposed modification allows to save the running time of the algorithms due to consideration of not all functional constraints on non-productive steps.

Remark 4 (Composite Optimization Problems [6,16,23]). Previously proposed methods are applicable to composite optimization problems, specifically

$$\min\{f(x) + r(x): \ x \in Q, \ g(x) + \eta(x) \leq 0\},$$

where $r, \eta : Q \to \mathbb{R}$ are so-called simple convex functionals (i.e. the proximal mapping operator $Mirr_h(x, \psi)$ is easily computable). For this case, for any $x, y \in Q$, we have

$$\psi_f(y, x) = \langle \nabla f(x), y - x \rangle + r(y) - r(x),$$
$$\psi_g(y, x) = \langle \nabla g(x), y - x \rangle + \eta(y) - \eta(x).$$

4   Stochastic Mirror Descent Algorithm

Let us, in this section, consider the stochastic setting of the problem (3). This means that we can still use the value of the objective function and functional constraints, but instead of their (sub)gradient, we use their stochastic (sub)gradient. Namely, we consider the first-order unbiased oracle that produces $\nabla f(x, \xi)$ and $\nabla g(x, \zeta)$, where ξ and ζ are random vectors and

$$\mathbb{E}[\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}[\nabla g(x, \zeta)] = \nabla g(x).$$

Assume that for each $x, y \in Q$

$$\langle \nabla f(x, \xi), x - y \rangle \leq M_f \sqrt{2V_d(y, x)} \text{ and } \langle \nabla g(x, \zeta), x - y \rangle \leq M_g \sqrt{2V_d(y, x)}, \quad (15)$$

where $M_f, M_g > 0$. Let us consider a proximal mapping operator for f

$$Mirr_h\left(x, \nabla f(x, \xi)\right) = \arg\min_{y \in Q} \left\{ \frac{1}{h} V_d(y, x) + \langle \nabla f(x, \xi), y \rangle \right\},$$

and, similarly, we consider a proximal mapping operator for g. The following lemma describes the main property of this operator.

Lemma 2. *Let f be a convex function which satisfies (4), $h > 0, \delta > 0, \xi$ be a random vector and $\tilde{x} = Mirr_h\left(x, \nabla f(x, \xi)\right)$. Then for all $y \in Q$*

$$h(f(x) - f(y)) \leq \phi_f^*(h) + V_d(y, x) - V_d(y, \tilde{x}) + h\langle \nabla f(x, \xi) - \nabla f(x), y - x \rangle + h\delta,$$

where, as earlier, $\phi_f^*(h) = \frac{h^2 M_f^2}{2}$.

Suppose $\varepsilon > 0$ is a given positive real number. We say that a (random) point $\widehat{x} \in Q$ is an expected ε–solution to the problem (3), in the stochastic setting, if

$$\mathbb{E}[f(\widehat{x})] - f(x^*) \leq \varepsilon \text{ and } g(\widehat{x}) \leq \varepsilon. \quad (16)$$

In order to solve the stochastic setting of the considered problem (3), we propose the following algorithm.

The following theorem gives information about the efficiency of the proposed Algorithm 4. The proof of this theorem is given in [29], Appendix B.

Algorithm 4. Modified Mirror Descent for the stochastic setting.

Require: $\varepsilon > 0, \delta > 0, h^f > 0, h^g > 0, \Theta_0 : d(x^*) \leq \Theta_0^2$.
1: $x^0 = \arg\min_{x \in Q} d(x)$.
2: $I =: \emptyset$ and $J =: \emptyset$
3: $N \leftarrow 0$
4: **repeat**
5:     **if** $g\left(x^N\right) \leq \varepsilon + \delta$ **then**
6:         $x^{N+1} = Mirr_{h^f}\left(x^N, \nabla f(x, \xi^N)\right),$ "productive step"
7:         $N \to I$
8:     **else**
9:         $x^{N+1} = Mirr_{h^g}\left(x^N, \nabla f(x, \zeta^N)\right),$ "non-productive step"
10:        $N \to J$
11:    **end if**
12:    $N \leftarrow N + 1$
13: **until** $\Theta_0^2 \leq \varepsilon\left(|J|h^g + |I|h^f\right) - |J|\phi_g^*(h^g) - |I|\phi_f^*(h^f)$.
Ensure: $\widehat{x} := \frac{1}{|I|} \sum\limits_{k \in I} x^k$.

Theorem 4. *Let f and g be convex functions and* (15) *hold. Let $\varepsilon > 0, \delta > 0$ be fixed positive numbers. Then, after the stopping of Algorithm 4, the following inequalities hold:*

$$\mathbb{E}[f(\widehat{x})] - f(x^*) \leq \varepsilon + \delta \quad and \quad g(\widehat{x}) \leq \varepsilon + \delta.$$

Remark 5. It should be noted how the optimality of the proposed method can be understood. With the special assumptions (10)–(11) and choice of h^f, h^g, the complexity of the algorithm is $O(\frac{1}{\varepsilon^2})$, which is optimal in such a class of problems.

5   Online Optimization Problem

In this section we consider the online setting of the optimization problem (3). Namely,

$$\frac{1}{N} \sum_{i=1}^{N} f_i(x) \to \min_{x \in Q, \, g(x) \leq 0}, \tag{17}$$

under the assumption that all $f_i : Q \to \mathbb{R}$ $(i = 1, \ldots, N)$ and g satisfy (10) and (11) with constants $M_i > 0, i = 1, \ldots, N$ and $M_g > 0$.

In order to solve problem (17), we propose an algorithm (listed as Algorithm 5 below). This algorithm produces N productive steps and in each step, the (sub)gradient of exactly one functional of the objectives is calculated. As a result of this algorithm, we get a sequence $\{x^k\}_{k \in I}$ (on productive steps), which can be considered as a solution to problem (17) with accuracy κ (see (18)).

Assume that $M = \max\{M_i, M_g\}, h^f = h^g = h = \frac{\varepsilon}{M}$.

Algorithm 5. Modified Mirror Descent for the online setting.

Require: $\varepsilon > 0, \delta > 0, M > 0, N, \Theta_0 : d(x^*) \leq \Theta_0^2$.
1: $x^0 = \arg\min_{x \in Q} d(x)$.
2: $i := 1, k := 0$
3: set $h = \frac{\varepsilon}{M^2}$
4: **repeat**
5:     **if** $g\left(x^k\right) \leq \varepsilon + \delta$ **then**
6:         $x^{k+1} = Mirr_h\left(x^k, \psi_{f_i}\right),$ "productive step"
7:         $i = i + 1,$
8:         $k = k + 1,$
9:     **else**
10:         $x^{k+1} = Mirr_h\left(x^k, \psi_g\right),$ "non-productive step"
11:         $k = k + 1,$
12:     **end if**
13: **until** $i = N + 1$.
14: Guaranteed accuracy:

$$\kappa = \frac{|J|}{N}\left(-\frac{\varepsilon}{2}\right) + \left(\frac{\varepsilon}{2} + \delta\right) + \frac{M^2\Theta_0^2}{N\varepsilon}. \tag{18}$$

For Algorithm 5, we have the following result.

Theorem 5. *Suppose all $f_i : Q \to \mathbb{R}$ $(i = 1, \ldots, N)$ and g satisfy* (10) *and* (11) *with constants $M_i > 0, i = 1, \ldots, N$ and $M_g > 0$, Algorithm 5 works exactly N productive steps. Then after the stopping of this Algorithm, the following inequality holds*

$$\frac{1}{N}\sum_{i=1}^{N} f_i(x^k) - \min_{x \in Q} \frac{1}{N}\sum_{i=1}^{N} f_i(x) \leq \kappa,$$

moreover, when the regret is non-negative, there will be no more than $O(N)$ non-productive steps.

The proof of this theorem is given in [29], Appendix C. In particular, note that the proposed method is optimal [13]: if for some $C > 0$, $\kappa \sim \varepsilon \sim \delta = \frac{C}{\sqrt{N}}$, then $|J| \sim O(N)$.

6   Numerical Experiments

To show the practical performance of the proposed Algorithms 2, 3 and their modified versions, which are listed as Algorithm 6 and Algorithm 7 in [29], in the case of many functional constraints, a series of numerical experiments were performed[1], for the well-known *Fermat-Torricelli-Steiner* problem, but with some non-smooth functional constraints.

---

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

For a given set $\{P_k = (p_{1k}, p_{2k}, \ldots, p_{nk}); k \in [r]\}$ of r points, in n-dimensional Euclidean space $\mathbb{R}^n$, we need to solve the considered optimization problem (3), where the objective function f is given by

$$f(x) := \frac{1}{r} \sum_{k=1}^{r} \sqrt{(x_1 - p_{1k})^2 + \ldots + (x_n - p_{nk})^2} = \frac{1}{r} \sum_{k=1}^{r} \|x - P_k\|_2. \qquad (19)$$

The functional constraint has the following form

$$g(x) = \max_{i \in [m]} \{g_i(x) = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \ldots + \alpha_{in} x_n\}. \qquad (20)$$

The coefficients $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}$, for all $i \in [m]$, in (20) and the coordinates of the points P_k, for all $k \in [r]$, are drawn from the normal (Gaussian) distribution with the location of the mode equaling 1 and the scale parameter equaling 2. We choose the standard Euclidean norm and the Euclidean distance function in $\mathbb{R}^n$, $\delta = 0$, starting point $x^0 = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$ and Q is the unit ball in $\mathbb{R}^n$.

We run Algorithms 2, 3 and their modified versions, Algorithms 6 and 7 respectively (see [29]), for $m = 200, n = 500, r = 100$ and different values of $\varepsilon \in \{\frac{1}{2^i} : i = 1, 2, 3, 4, 5\}$. The results of the work of these algorithms are represented in Table 1 below. These results demonstrate the comparison of the number of iterations (Iter.), the running time (in seconds) of each algorithm and the qualities of the solution, produced by these algorithms with respect to the objective function f and the functional constraint g, where we calculate the values of these functions at the output $x^{\text{out}} := \hat{x}$ of the algorithms. We set $f^{\text{best}} := f(x^{\text{out}})$ and $g^{\text{out}} := g(x^{\text{out}})$.

Table 1. The results of Algorithms 2, 3 and their modified versions Algorithms 6 and 7 respectively, with $m = 200, n = 500, r = 100$ and different values of ε.

	Algorithm 2				Algorithm 6			
$1/\varepsilon$	Iter	Time (sec.)	f^{best}	g^{out}	Iter	Time (sec.)	f^{best}	g^{out}
2	16	5.138	22.327427	2.210041	16	4.883	22.327427	2.210041
4	64	20.911	22.303430	2.016617	64	20.380	22.303430	2.016617
8	256	84.343	22.283362	1.858965	256	79.907	22.283362	2.015076
16	1024	317.991	22.274366	1.199792	1024	317.033	22.273177	1.988190
32	4096	1253.717	22.272859	0.607871	4096	1145.033	22.269038	1.858965
	Algorithm 3				Algorithm 7			
2	167	9.455	22.325994	0.417002	164	7.373	22.325604	0.391461
4	710	39.797	22.305980	0.204158	667	29.954	22.305654	0.188497
8	2910	158.763	22.289320	0.103493	2583	119.055	22.289302	0.088221
16	11613	626.894	22.280893	0.051662	10155	468.649	22.280909	0.045343
32	46380	2511.261	22.277439	0.026000	40149	1723.136	22.277450	0.022639

In general, from the conducted experiments, we can see that Algorithm 2 and its modified version (Algorithm 6) work faster than Algorithms 3 and its modified version (Algorithm 7). But note that Algorithms 3 and 7 guarantee a better quality of the resulting solution to the considered problem, with respect to the objective function f and the functional constraint (20). Also, we can see the efficiency of the modified Algorithm 7, which saves the running time of the algorithm, due to consideration of not all functional constraints on non-productive steps.

7 Conclusion

In the paper, there was introduced the concept of an inexact (δ, ϕ, V)–model of the objective function. There were considered some modifications of the Mirror Descent algorithm, in particular for stochastic and online optimization problems. A significant part of the work was devoted to the research of a special case of relative Lipschitz condition for the objective function and functional constraints. The proposed methods are applicable for a wide class of problems because relative Lipschitz-continuity is an essential generalization of the classical Lipschitz-continuity. However, for relatively Lipschitz-continuous problems, we could not propose adaptive methods like [3,27,28]. Note that Algorithm 3 and its modified version Algorithm 7 (see [29]) are partially adaptive since the resulting number of iterations is not fixed, due to the stopping criterion, although the step-sizes are fixed.

References

1. Alkousa, M.S.: On some stochastic mirror descent methods for constrained online optimization problems. Comput. Res. Model. **11**(2), 205–217 (2019)
2. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma Beyond Lipschitz Gradient continuity: first-order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2017)
3. 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
4. Beck, A., Ben-Tal, A., Guttmann-Beck, N., Tetruashvili, L.: The CoMirror algorithm for solving nonsmooth constrained convex problems. Oper. Res. Lett. **38**(6), 493–498 (2010)
5. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)
6. Beck, A., Teboulle, M.: A fast iterative Shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
7. Belmega, E.V., Mertikopoulos, P.: Online and stochastic optimization beyond Lipschitz continuity: a Riemannian approach. Published as a Conference Paper at ICLR (2020). https://openreview.net/pdf?id=rkxZyaNtwB

8. Devolder, O., Glineur, F., Nesterov, Yu.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1), 37–75 (2013). https://doi.org/10.1007/s10107-013-0677-5

9. Doan, T.T., Bose, S., Nguyen, D.H., Beck, C.L.: Convergence of the iterates in mirror descent methods. IEEE Control Syst. Lett. **3**(1), 114–119 (2019)

10. 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

11. Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. Autom. Remote Control **78**(2), 224–234 (2017). https://doi.org/10.1134/S0005117917020035

12. Hao, Y., Neely, M.J., Xiaohan, W.: Online convex optimization with stochastic constraints. In: Published in NIPS, pp. 1427–1437 (2017)

13. Hazan, E., Kale, S.: Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. JMLR **15**, 2489–2512 (2014)

14. Hazan, E.: Introduction to online convex optimization. Found. Trends Optim. **2**(3–4), 157–325 (2015)

15. Jenatton, R., Huang, J., Archambeau, C.: Adaptive algorithms for online convex optimization with long-term constraints. In: Proceedings of the 33rd International Conference on Machine Learning, PMLR 48, pp. 402–411 (2016)

16. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. **28**(1), 333–354 (2018)

17. Lu, H.: Relative continuity for Non-Lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. INFORMS J. Optim. **1**(4), 288–303 (2019)

18. Luong, D.V.N., Parpas, P., Rueckert, D., Rustem, B.: A weighted mirror descent algorithm for nonsmooth convex optimization problem. J. Optim. Theory Appl. **170**(3), 900–915 (2016). https://doi.org/10.1007/s10957-016-0963-5

19. Nazin, A.V., Miller, B.M.: Mirror descent algorithm for homogeneous finite controlled Markov Chains with unknown mean losses. In: IFAC Proceedings, vol. 44(1), pp. 12421–12426 (2011). https://doi.org/10.3182/20110828-6-IT-1002.03450

20. Nazin, A., Anulova, S., Tremba, A.: Application of the mirror descent method to minimize average losses coming by a poisson flow. In: 2014 European Control Conference (ECC), Strasbourg, pp. 2194–2197 (2014)

21. Nemirovskii, A.: Efficient Methods for Large-Scale Convex Optimization Problems. Math. Ekon. (1979). (in Russian)

22. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)

23. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**, 125–161 (2013)

24. Nesterov, Y.: Relative smoothness: new paradigm in convex optimization. In: Conference report, EUSIPCO-2019, A Coruna, Spain, 4 September 2019 (2019). https://www.gtec.udc.es/eusipco2019/wp-content/uploads/2019/10/Relative-Smoothness-New-Paradigm-in-Convex.pdf

25. Orabona, F.: A modern introduction to online learning (2020). https://arxiv.org/pdf/1912.13213.pdf

26. Orabona, F., Crammer, K., Cesa-Bianchi, N.: A generalized online mirror descent with applications to classification and regression. Mach. Learn. **99**(3), 411–435 (2014). https://doi.org/10.1007/s10994-014-5474-8

27. Stonyakin, F.S., Alkousa, M., Stepanov, A.N., Titov, A.A.: Adaptive mirror descent algorithms for convex and strongly convex optimization problems with functional constraints. J. Appl. Ind. Math. **13**(3), 557–574 (2019). https://doi.org/10.1134/S1990478919030165

28. Stonyakin, F.S., Stepanov, A.N., Gasnikov, A.V., Titov, A.A.: Mirror descent for constrained optimization problems with large subgradient values of functional constraints. Comput. Res. Model. **12**(2), 301–317 (2020)

29. Titov, A.A., Stonyakin, F.S., Alkousa, M.S., Ablaev, S.S., Gasnikov, A.V.: Analogues of switching subgradient schemes for relatively Lipschitz-continuous convex programming problems (2020). https://arxiv.org/pdf/2003.09147.pdf

30. Titov, A.A., Stonyakin, F.S., Gasnikov, A.V., Alkousa, M.S.: Mirror descent and constrained online optimization problems. In: Evtushenko, Y., Jaćimović, M., Khachay, M., Kochetov, Y., Malkova, V., Posypkin, M. (eds.) OPTIMA 2018. CCIS, vol. 974, pp. 64–78. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10934-9_5

31. Tremba, A., Nazin, A.: Extension of a saddle point mirror descent algorithm with application to robust PageRank. In: 52nd IEEE Conference on Decision and Control, 10–13 December, pp. 3691–3696 (2013)

32. Tyurin, A.I., Gasnikov, A.V.: Fast gradient descent method for convex optimization problems with an oracle that generates a (δ, L)-model of a function in a requested point. Comput. Math. Math. Phys. **59**(7), 1137–1150 (2019)

33. Yuan, J., Lamperski, A.: Online convex optimization for cumulative constraints. In: Published in NIPS, pp. 6140–6149 (2018)

34. Yunwen, L., Ding-Xuan, Z.: Convergence of online mirror descent. Appl. Comput. Harmon. Anal. **48**(1), 343–373 (2020)

35. Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S.P., Glynn, P.W.: On the convergence of mirror descent beyond stochastic convex programming. SIAM J. Optim. **30**(1), 687–716 (2020)