

Online Optimization Problems with Functional Constraints under Relative Lipschitz Continuity and Relative Strong Convexity Conditions*

Oleg Savchuk^{1,2}[0000-0003-3732-1855], Fedor Stonyakin^{1,2}[0000-0002-9250-4438],
 Mohammad Alkousa^{1,3}[0000-0001-5470-0182], Rida
 Zabirowa¹[0000-0000-0000-0000], Alexander Titov¹[0000-0001-9672-0616], and
 Alexander Gasnikov^{1,3,4,5}[0000-0002-5982-8983]

¹ Moscow Institute of Physics and Technology, 9 Institutsky lane, Dolgoprudny, 141701, Russia

² V.I. Vernadsky Crimean Federal University, 4 Academician Vernadsky Avenue, Simferopol, 295007, Republic of Crimea, Russia

³ HSE University, Moscow, 20 Myasnitskaya street, Moscow, 101000, Russia

⁴ Institute for Information Transmission Problems RAS, 11 Pokrovsky boulevard, 109028, Moscow, Russia

⁵ Caucasus Mathematical Center, Adyghe State University, 208 Pervomaiskaya street, Maykop, Republic of Adygea, 385000, Russia

oleg.savchuk19@mail.ru, fedyor@mail.ru, mohammad.alkousa@phystech.edu, a.a.tytov@gmail.com, zabirowa.rr@phystech.edu, gasnikov.av@phystech.edu

Abstract. A few years ago, the optimization field introduced classes of relatively smooth [2], relatively continuous, and relatively strongly convex optimization problems [5,10]. These concepts have expanded the class of problems to which optimal complexity estimates of gradient-type methods in high-dimensional spaces can be applied. There are known works on online optimization (regret minimization) problems for both relatively Lipschitz and relatively strongly convex problems. In this work, we consider the problem of strongly convex online optimization with convex inequality constraints. A scheme with switching over productive and non-productive steps is proposed for these problems. The convergence rate of the proposed scheme is proven for the class of relatively Lipschitz and strongly convex minimization problems. Moreover, analogously with the [6] we study extensions of the considered Mirror Descent algorithms that eliminate the need for a priori knowledge of the lower bound on the (relative) strong convexity parameters of the observed functions. Some numerical experiments were conducted to demonstrate the effectiveness of one of the proposed algorithms with a comparison with another adaptive algorithm for convex online-optimization problems.

Keywords: Online Optimization · Strongly Convex Programming Problem · Relatively Lipschitz-Continuous Function · Relatively Strongly Convex Function · Mirror Descent · Regularization.

* The research was supported by Russian Science Foundation (project No. 21-71-30005), <https://rscf.ru/en/project/21-71-30005/>.

Introduction

The development of numerical methods for solving non-smooth online optimization problems presents a great interest nowadays due to the appearance of many applied problems with the corresponding statement [3,6,7,8,11]. Online optimization plays a key role in solving machine learning, finance, networks, and other problems. As some examples of such problems, we can mention multi-armed bandits, job-shop scheduling and ski rental problems, search games, etc. One of the most popular methods of solving online optimization problems is the Mirror Descent method [14]. Let us note, that Mirror Descent can be also applied for solving online optimization problems in a stochastic setting [1,4], which allows using an arbitrary, not necessarily 1-strongly convex, distance-generating function (see (6)).

Remind, that the online optimization problem represents the problem of minimizing the sum (or the arithmetic mean) of T functionals $f_t : Q \rightarrow \mathbb{R}$ ($t = \overline{1, T}$) given on some closed convex set $Q \subset \mathbb{R}^n$

$$\min_{x \in Q} \frac{1}{T} \sum_{t=1}^T f_t(x), \quad s.t. \quad g(x) \leq 0. \quad (1)$$

The key feature of the problem statement consists in the possibility of calculating the (sub)gradient $\nabla f_t(x)$ of each functional f_t only once.

Recently, in [16] there were proposed some modifications of the Mirror Descent method for solving online optimization problems in the case, if all the functions $f_t(x)$ and functional constraint $g(x)$ satisfy Lipschitz condition, i.e. there exists such a constant $M > 0$, that

$$|g(x) - g(y)| \leq M \|x - y\|, \quad (2)$$

$$|f_t(x) - f_t(y)| \leq M \|x - y\|, \quad \forall t = \overline{1, T}. \quad (3)$$

In the case of non-negativity of regret

$$Regret_T := \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x), \quad (4)$$

these methods are optimal for the considered class of problems accordingly to [7], the number of non-productive steps during their work is $O(T)$. In the case of negative regret, the number of non-productive steps for the proposed methods is $O(T^2)$.

Later, in [17] the smoothness class for the applicability of such approaches has been extended by reducing the requirement of Lipschitz continuity of functions to the recently proposed concept of relative Lipschitz continuity [9,12].

Definition 1. *Let us call a convex function $f : Q \rightarrow \mathbb{R}$ M -relatively Lipschitz-continuous for some $M > 0$, if the following inequality holds*

$$\langle \nabla f(x), y - x \rangle + M \sqrt{2V(y, x)} \geq 0, \quad \forall x, y \in Q. \quad (5)$$

This concept has been widely used in many applied problems and has also enabled the proposal of subgradient methods for both non-differentiable and non-Lipschitz Support Vector Machine (SVM) and for problems of Intersection of n Ellipsoids while maintaining optimal convergence rate estimates for the class of simply Lipschitz-continuous functions. It is worth noting that the proposed method also allowed the use of an imprecisely defined function (more exactly, a function that admits a representation in a model form), nevertheless, the method was also optimal.

Let $h : Q \rightarrow \mathbb{R}$ be a distance-generating function (or prox-function) that is continuously differentiable and convex. For all $x, y \in Q$ we consider the corresponding Bregman divergence

$$V(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle. \quad (6)$$

In this paper, we improve existing estimates of the convergence rate by considering a class of strongly convex functions and generalize the obtained problem statement to the case of problems with functional constraints.

Definition 2. *A function f over a convex set Q is called μ -strongly convex with respect to a convex function h if*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \mu V(x, y), \quad \forall x, y \in Q,$$

More precisely, we present a novel theorem that provides a tighter bound on regret, in terms of the number of productive steps taken by the algorithm. Specifically, the theorem proves that if the algorithm completes exactly T productive steps and has a non-negative regret, then the number of non-productive steps satisfies $T_j \leq CT$, where C is a constant. This result significantly improves existing convergence rate estimates for the Mirror Descent method with functional constraints. In addition, we obtain the complexity of the bound in terms of T and some other problem parameters. This corollary allows us to determine the number of productive steps needed to achieve the desired accuracy of regret in practice.

We also consider some modifications of the Mirror Descent method for solving non-smooth online optimization problems [6]. Specifically, the paper introduces two algorithms for solving strongly convex minimization problems with and without regularization. The first algorithm, called General-Norm Online Gradient Descent: Relatively Strongly Convex and Relatively Lipschitz-Continuous Case, is based on a convex function h and updates the solution iteratively using predictions and observations of the objective function f_t . The second algorithm, called Adaptive General-Norm Online Gradient Descent with Regularization, extends the first algorithm by introducing an adaptive regularization term that depends on a function d that is both relatively Lipschitz continuous and relatively strongly convex.

For each algorithm, we provide the theoretical justification of bounds on the regret. These theorems guarantee upper bounds on the regret for each algorithm and can be used to analyze the performance of the algorithms. Overall, the paper

presents a comprehensive framework for solving non-smooth online optimization problems with functional constraints, and the results have practical implications for a broad range of applications.

The paper consists of an introduction and 4 main sections. In Sect. 1 we consider the basic statement of the constrained online optimization problem and propose a modification of the Mirror Descent method for minimizing the arithmetic mean of relatively strongly convex and relatively Lipschitz-continuous functionals, supposing that functional constraint satisfies the same conditions. We also provide a theoretical justification for the convergence rate of the proposed method. Sect. 2 is devoted to some modifications of the algorithms, proposed in [6] for the corresponding class of problems with regularization. In Sect. 3 we combine the above-mentioned ideas and propose algorithms with switching over productive and non-productive steps both with and without iterative regularization during the work of algorithms. In Sect. 4 we present some numerical experiments which demonstrate the effectiveness of one of the proposed algorithms and a comparison with another adaptive algorithm for the considered optimization problems.

To sum it up, the contributions of the paper can be stated as follows:

- We proposed an optimal method for solving a constrained online optimization problem with relatively strongly convex and relatively Lipschitz-continuous objective functionals and functional constraints. For the case of non-negative regret, the number of non-productive steps is bounded by $O(T)$.
- We proposed two algorithms for solving strongly convex minimization problems with and without regularization based on iteratively updating steps by using some auxiliary functions. Similar to [6], we present extensions of Mirror Descent that exclude the need for a priori knowledge of the lower bound on the (relatively) strong convexity parameters of the observed functions.
- We provided the results of numerical experiments demonstrating the advantages of using the proposed methods.

1 Mirror Descent for Relatively Strongly Convex and Relatively Lipschitz-Continuous Online-optimization Problems with Inequality Constraints

In this section, we present a scheme with switching over productive and non-productive steps for relatively strongly convex and relatively Lipschitz-continuous online optimization problems with inequality constraints. We consider the following strongly convex constrained optimization problem

$$\min_{x \in Q} \sum_{t=1}^T f_t(x), \quad g(x) \leq 0, \quad (7)$$

where $f_t : Q \rightarrow \mathbb{R}$ and $g : Q \rightarrow \mathbb{R}$. Let x^* be a solution of (7), i.e.

$$x^* = \arg \min_{x \in Q} \sum_{t=1}^T f_t(x), \quad g(x^*) \leq 0.$$

Let us denote the set of productive steps x_t for which $g(x_t) \leq \varepsilon$ by I , and the set of non-productive steps by J . Let $T = |I|, T_J = |J|$. Let us consider a subgradient method with switching over productive and non-productive steps.

Algorithm 1 Constrained Online Optimization: Mirror Descent for Relatively Lipschitz-Continuous and Relatively Strongly Convex Problems.

Require: $\varepsilon > 0, \mu > 0, T, x_1 \in Q$.

```

1:  $i := 1, t := 1$ ;
2: repeat
3:   if  $g(x_t) \leq \varepsilon$  then
4:      $\eta_t = \frac{1}{\mu t}$ ;
5:      $x_{t+1} := \text{Pr}_Q\{x_t - \eta_t \nabla f_t(x_t)\}$ ;   "productive step"
6:      $i := i + 1$ ;
7:      $t := t + 1$ ;
8:   else
9:      $\eta_t = \frac{1}{\mu t}$ ;
10:     $x_{t+1} := \text{Pr}_Q\{x_t - \eta_t \nabla g(x_t)\}$ ;   "non-productive step"
11:     $t := t + 1$ ;
12:   end if
13: until  $i = T + 1$ .
    
```

Theorem 1. *Suppose that, for each t , f_t is an M_f -relatively Lipschitz continuous and μ -strongly convex function with respect to the prox-function h . Let $g(x)$ be M_g -relatively Lipschitz continuous and μ -strongly convex function with respect to h . Suppose that Algorithm 1 for*

$$\varepsilon = \frac{M^2}{\mu} \frac{1 + \ln T}{T}$$

where $M = \max\{M_f, M_g\}$, works exactly T productive steps and $\text{Regret}_T \geq 0$. Then there exists a constant $C \in (2, 3)$ such that the number of non-productive steps satisfies $T_J \leq CT$, moreover, the following inequality holds:

$$\text{Regret}_T := \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \leq \frac{M^2}{\mu} \left(1 + \ln((C+1)T) \right) = O(T\varepsilon),$$

where $g(x_t) \leq \varepsilon$ for any $t = \overline{1, T}$.

Proof. Let us check the auxiliary inequality

$$\sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \leq \frac{M^2}{\mu} (1 + \ln(T + T_J)) - \varepsilon T_J. \quad (8)$$

1. Taking into account the M_f -relative Lipschitz-continuity of the function f_t for each productive step we have

$$\begin{aligned} \eta_t \left(f_t(x_t) - f_t(x^*) \right) &\leq \eta_t \left(\langle \nabla f_t, x_t - x^* \rangle - \mu V(x^*, x_t) \right) \\ &\leq \eta_t^2 M_f^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t \mu V(x^*, x_t). \end{aligned}$$

Hence, after dividing both sides of the above inequality by η_t we get

$$\begin{aligned} f_t(x_t) - f_t(x^*) &\leq \eta_t M_f^2 + \frac{1}{\eta_t} \left(V(x^*, x_t) - V(x^*, x_{t+1}) \right) - \mu V(x^*, x_t) \\ &= \frac{M_f^2}{\mu t} + \mu t V(x^*, x_t) - \mu V(x^*, x_t) - \mu t V(x^*, x_{t+1}) \quad (9) \\ &= \frac{M_f^2}{\mu t} + \mu(t-1)V(x^*, x_t) - \mu t V(x^*, x_{t+1}). \end{aligned}$$

2. Similarly, taking into account the M_g -relative Lipschitz-continuity of g for each non-productive step we have $g(x_t) > \varepsilon$ and

$$\begin{aligned} \eta_t \varepsilon &< \eta_t (g(x_t) - g(x^*)) \leq \eta_t \left(\langle \nabla g, x_t - x^* \rangle - \mu V(x^*, x_t) \right) \\ &\leq \eta_t^2 M_g^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t \mu V(x^*, x_t). \end{aligned}$$

Dividing both sides of the last inequality by η_t , we get:

$$\begin{aligned} \varepsilon &< g(x_t) - g(x^*) \\ &\leq \eta_t M_g^2 + \frac{1}{\eta_t} \left(V(x^*, x_t) - V(x^*, x_{t+1}) \right) - \mu V(x^*, x_t) \\ &= \frac{M_g^2}{\mu t} + \mu t V(x^*, x_t) - \mu V(x^*, x_t) - \mu t V(x^*, x_{t+1}) \quad (10) \\ &= \frac{M_g^2}{\mu t} + \mu(t-1)V(x^*, x_t) - \mu t V(x^*, x_{t+1}). \end{aligned}$$

3. Summing up inequalities (9), (10) over productive and non-productive steps, for $M = \max\{M_f, M_g\}$, we get

$$\begin{aligned} &\sum_{t \in I} \left(f_t(x_t) - f_t(x^*) \right) + \sum_{t \in J} \left(g(x_t) - g(x^*) \right) \\ &\leq \sum_{t=1}^{T+T_J} \left(\frac{M^2}{\mu t} + \mu(t-1)V(x^*, x_t) - \mu t V(x^*, x_{t+1}) \right) \\ &\leq \frac{M^2}{\mu} \ln(T + T_J) - \mu(T + T_J)V(x^*, x_{T+T_J}) \\ &\leq \frac{M^2}{\mu} \ln(T + T_J). \end{aligned}$$

Using the fact, that for non-productive steps

$$g(x_t) - g(x^*) \geq g(x_t) > \varepsilon,$$

we get an estimate for the sum of the objective functionals:

$$\begin{aligned} \sum_{t \in I} (f_t(x_t) - f_t(x^*)) &\leq \frac{M^2}{\mu} \ln(T + T_J) - \sum_{t \in J} (g(x_t) - g(x^*)) \\ &\leq \frac{M^2}{\mu} \ln(T + T_J) - \sum_{t \in J} \varepsilon = \frac{M^2}{\mu} \ln(T + T_J) - \varepsilon T_J. \end{aligned}$$

4. According to the assumption of non-negativity of the regret, we find

$$\begin{aligned} 0 \leq \text{Regret}_T &= \sum_{t=1}^T (f_t(x_t) - f_t(x^*)) = \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \\ &\leq \frac{M^2}{\mu} (1 + \ln(T + T_J)) - \varepsilon T_J. \end{aligned}$$

Hence $\varepsilon T_J \leq \frac{M^2}{\mu} (1 + \ln(T + T_J))$ and $\varepsilon = \frac{M^2}{\mu} \frac{1 + \ln T}{T}$. Therefore, we have

$$\begin{aligned} \frac{1 + \ln T}{T} T_J &\leq 1 + \ln(T + T_J), \\ \frac{T_J}{T} &\leq \frac{1 + \ln(T + T_J)}{1 + \ln T}. \end{aligned}$$

Moreover, taking into account

$$\ln(T + T_J) = \ln \left(T \left(1 + \frac{T_J}{T} \right) \right) = \ln T + \ln \left(1 + \frac{T_J}{T} \right),$$

we get

$$\frac{T_J}{T} \leq \frac{1 + \ln T + \ln(1 + \frac{T_J}{T})}{1 + \ln T} \leq 1 + \ln \left(1 + \frac{T_J}{T} \right).$$

Since the linear function grows faster than the logarithmic one, it is obvious, that for a sufficiently large T_J , the above inequality does not hold, thus $\frac{T_J}{T}$ is bounded. Therefore, we proved that $T_J = O(T)$, i.e. there exists $C > 0$ such that $T_J \leq CT$ or $\frac{T_J}{T} \leq C$:

$$\frac{T_J}{T} \leq 1 + \ln \left(1 + \frac{T_J}{T} \right).$$

Equality in the latter inequality is achieved when

$$\frac{T_J}{T} \approx 2, 146.$$

5. Further, we note that by the definition of ε , we have

$$\varepsilon = \frac{M^2}{\mu} \frac{1 + \ln T}{T} = \frac{M^2}{\mu T} + \frac{M^2}{\mu} \frac{\ln T}{T}.$$

Since T is the number of productive steps and $T_J \leq CT$ is the number of non-productive steps, the total number of steps is $T + T_J \leq (C + 1)T$. Therefore

$$\varepsilon = \frac{M^2}{\mu(T + T_J)} + \frac{M^2 \ln(T + T_J)}{\mu(T + T_J)} \leq \frac{M^2}{\mu(C + 1)T} + \frac{M^2 \ln(C + 1)T}{\mu(C + 1)T}.$$

This allows us to bound the regret as follows:

$$\text{Regret}_T := \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \leq \frac{M^2}{\mu} (1 + \ln(C + 1)T).$$

This shows that the bound on the regret, given by the last inequality holds, which finishes the proof.

Remark 1. Let us show that our algorithm will necessarily make at least one productive steps. Indeed, suppose, that the number of productive steps equals zero, then

$$\varepsilon T_J \leq \sum_{t=1}^{T_J} (g(x_t) - g(x^*)) \leq \frac{M^2}{\mu} (1 + \ln T_J).$$

It is obviously, that for a sufficiently large T_J , the above inequality does not hold. Thus, for a sufficiently large number of non-productive steps, there will be at least one productive step.

Let us find out how many non-productive steps need to be taken to achieve inequality:

$$\begin{aligned} \varepsilon T_J &= \frac{T_J M^2}{\mu} \frac{1 + \ln T}{T} > \frac{M^2}{\mu} (1 + \ln T_J), \\ \frac{1 + \ln T}{T} &> \frac{1 + \ln T_J}{T_J}. \end{aligned}$$

Then $T_J \leq CT$, where C is a constant, which proves that the number of non-productive steps is bounded until at least one productive step is made.

2 Online Mirror Descent with Regularization

In this section, we propose some modifications of the algorithms proposed in [6] for relatively strongly convex and relatively Lipschitz online optimization problems and provide theoretical estimates of the quality of the solution.

We consider the following strongly convex minimization problem

$$\min_{x \in Q} \sum_{t=1}^T f_t(x), \tag{11}$$

where $f_t : Q \rightarrow \mathbb{R}$. Define $\mu_{1:t} := \sum_{s=1}^t \mu_s$, where μ_s is the parameter of relative strong convexity of the function f_s . Let $\mu_{1:0} = 0$.

Algorithm 2 General-Norm Online Gradient Descent: Relatively Strongly Convex and Relatively Lipschitz-Continuous Case.

- 1: Input: convex function h .
 - 2: Initialize x_1 arbitrarily.
 - 3: **for** $t = 1, \dots, T$ **do**
 - 4: Predict x_t , observe f_t .
 - 5: Compute η_{t+1} and let y_{t+1} be such that $\nabla h(y_{t+1}) = \nabla h(x_t) - \eta_{t+1} \nabla f_t(x_t)$.
 - 6: Let $x_{t+1} = \arg \min_{x \in Q} V(x, y_{t+1})$ be the projection of y_{t+1} onto Q .
 - 7: **end for**
-

Theorem 2. *Suppose that, for each t , f_t is an M_t -relatively Lipschitz-continuous and μ_t -strongly convex function with respect to prox-function h . Applying the Algorithm 2 with $\eta_{t+1} = \frac{1}{\mu_{1:t}}$, we have*

$$\text{Regret}_T \leq \sum_{t=1}^T \frac{M_t^2}{\mu_{1:t}}.$$

Proof. The proof is given in Appendix A.

Let's now consider an analogue of Algorithm 2 for relatively strongly convex and relatively Lipschitz-continuous problems with iterative regularization. Define

$\lambda_{1:t} := \sum_{s=1}^t \lambda_s$. The proposed algorithm is listed as Algorithm 3, below.

Algorithm 3 Adaptive General-Norm Online Gradient Descent with Regularization.

- 1: Input: convex function h .
- 2: Initialize x_1 arbitrarily.
- 3: **for** $t = 1, \dots, T$ **do**
- 4: Predict x_t , observe f_t .
- 5: Compute $\lambda_t = \frac{1}{2} \left(\sqrt{(\mu_{1:t} + \lambda_{1:t-1})^2 + 8M_t^2 / (A^2 + 2M_d^2)} - (\mu_{1:t} + \lambda_{1:t-1}) \right)$.
- 6: Compute η_{t+1} and let y_{t+1} be such that

$$\nabla h(y_{t+1}) = \nabla h(x_t) - \eta_{t+1} (\nabla f_t(x_t) + \lambda_t \nabla d(x_t)).$$

- 7: Let $x_{t+1} = \arg \min_{x \in Q} V(x, y_{t+1})$ be the projection of y_{t+1} onto Q .
 - 8: **end for**
-

For Algorithm 3, we have the following result.

Theorem 3. *Suppose that, for each t , f_t is M_t -relatively Lipschitz-continuous and μ_t -relatively strongly convex function with respect to the prox-function h . Let $d : Q \rightarrow \mathbb{R}$ be M_d -relatively Lipschitz-continuous and 1-strongly convex*

function with respect to h . Suppose that $d(x) \geq 0$, $\forall x \in Q$ and $A^2 = \sup_{x \in Q} d(x)$. Applying Algorithm 3 with $\eta_{t+1} = \frac{1}{\mu_{1:t} + \lambda_{1:t}}$, the following inequalities hold

$$\text{Regret}_T \leq \lambda_{1:T} A^2 + \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}},$$

and

$$\text{Regret}_T \leq 2 \inf_{\lambda_1^*, \dots, \lambda_T^*} \left((A^2 + 2M_d^2) \lambda_{1:T}^* + \sum_{t=1}^T \frac{(M_t + \lambda_t^* M_d)^2}{\mu_{1:t} + \lambda_{1:t}^*} \right).$$

Proof. The proof is given in Appendix B.

3 The Case of Online Optimization Problems with Functional Constraints

In this section, we consider a scheme with switching over productive and non-productive steps both with and without iterative regularization for a relatively strongly convex and relatively Lipschitz-continuous constrained online optimization problem.

Remind that we consider the following problem of strongly convex conditional minimization

$$\min_{x \in Q} \sum_{t=1}^T f_t(x), \quad g(x) \leq 0.$$

and

$$x^* = \arg \min_{x \in Q} \sum_{t=1}^T f_t(x), \quad g(x^*) \leq 0,$$

where $f_t : Q \rightarrow \mathbb{R}$ and $g : Q \rightarrow \mathbb{R}$. Remind that the set of productive steps is I , the set of non-productive steps is J and $T = |I|, T_J = |J|$. Similarly to Section 2, we define $\mu_{1:t} := \sum_{s=1}^t \mu_s$, where μ_s is the parameter of relative strong convexity of the function f_s and let $\mu_{1:0} = 0$. If t is the number of non-productive step, then $\mu_t = \mu_g$, where μ_g is the parameter of relative strong convexity of the function g .

Theorem 4. *Suppose that, for each t , f_t is an M_t -relatively Lipschitz-continuous and μ_t -strongly convex function with respect to the convex function h . Let $g(x)$ be M_g -relatively Lipschitz-continuous and μ_g -strongly convex function with respect to h . If Algorithm 4 works exactly T productive steps and $\text{Regret}_T \geq 0$, then the following inequality holds:*

$$\text{Regret}_T \leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \varepsilon T_J,$$

where $M = \max\{M_t, M_g\}$ and $g(x_t) \leq \varepsilon$ for any $t = \overline{1, T}$.

Algorithm 4 Mirror Descent for Constrained Optimization Problems with Relatively Lipschitz-Continuous and Relatively Strongly Convex Functions.

Require: $\varepsilon > 0, T, x_1 \in Q$.

- 1: $i := 1, t := 1$;
- 2: **repeat**
- 3: **if** $g(x_t) \leq \varepsilon$ **then**
- 4: $\eta_t = \frac{1}{\mu_{1:t}}$;
- 5: $x_{t+1} := \text{Pr}_Q\{x_t - \eta_t \nabla f_t(x_t)\}$; "productive step"
- 6: $i := i + 1$;
- 7: $t := t + 1$;
- 8: **else**
- 9: $\eta_t = \frac{1}{\mu_{1:t}}$;
- 10: $x_{t+1} := \text{Pr}_Q\{x_t - \eta_t \nabla g(x_t)\}$; "non-productive step"
- 11: $t := t + 1$;
- 12: **end if**
- 13: **until** $i = T + 1$.
- 14: Guaranteed accuracy:

$$\delta := \frac{1}{T} \left(\sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \varepsilon T_J \right).$$

Proof. 1. Taking into account that f_t is M_t -relative Lipschitz continuous, then for every productive step, we have

$$\begin{aligned} \eta_t (f_t(x_t) - f_t(x^*)) &\leq \eta_t (\langle \nabla f_t, x_t - x^* \rangle - \mu_t V(x^*, x_t)) \\ &\leq \eta_t^2 M_t^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t \mu_t V(x^*, x_t). \end{aligned}$$

Hence, after dividing both sides of the above inequality by η_t , we get

$$\begin{aligned} f_t(x_t) - f_t(x^*) &\leq \eta_t M_t^2 + \frac{1}{\eta_t} (V(x^*, x_t) - V(x^*, x_{t+1})) - \mu_t V(x^*, x_t) \\ &= \frac{M_t^2}{\mu_{1:t}} + \mu_{1:t} V(x^*, x_t) - \mu_t V(x^*, x_t) - \mu_{1:t} V(x^*, x_{t+1}) \\ &= \frac{M_t^2}{\mu_{1:t}} + \mu_{1:t-1} V(x^*, x_t) - \mu_{1:t} V(x^*, x_{t+1}). \end{aligned} \tag{12}$$

2. Similarly, taking into account that g is M_g -relative Lipschitz continuous, then for every non-productive step, we have $g(x_t) > \varepsilon$, and

$$\begin{aligned} \eta_t \varepsilon < \eta_t (g(x_t) - g(x^*)) &\leq \eta_t (\langle \nabla g, x_t - x^* \rangle - \mu_t V(x^*, x_t)) \\ &\leq \eta_t^2 M_g^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t \mu_t V(x^*, x_t). \end{aligned}$$

Dividing both sides of the last inequality by η_t , we get:

$$\begin{aligned}
\varepsilon &< g(x_t) - g(x^*) \\
&\leq \eta_t M_g^2 + \frac{1}{\eta_t} \left(V(x^*, x_t) - V(x^*, x_{t+1}) \right) - \mu_t V(x^*, x_t) \\
&= \frac{M_g^2}{\mu_{1:t}} + \mu_{1:t} V(x^*, x_t) - \mu_t V(x^*, x_t) - \mu_{1:t} V(x^*, x_{t+1}) \\
&= \frac{M_g^2}{\mu_{1:t}} + \mu_{1:t-1} V(x^*, x_t) - \mu_{1:t} V(x^*, x_{t+1}).
\end{aligned} \tag{13}$$

3. Summing up inequalities (12), (13) over productive and non-productive steps, for $M = \max\{M_t, M_g\}$, we get

$$\begin{aligned}
&\sum_{t \in I} \left(f_t(x_t) - f_t(x^*) \right) + \sum_{t \in J} \left(g(x_t) - g(x^*) \right) \\
&\leq \sum_{t=1}^{T+T_J} \left(\frac{M^2}{\mu_{1:t}} + \mu_{1:t-1} V(x^*, x_t) - \mu_{1:t} V(x^*, x_{t+1}) \right) \\
&\leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \mu_{1:T+T_J} V(x^*, x_{T+T_J}) \leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}}.
\end{aligned}$$

Using the fact, that for non-productive steps

$$g(x_t) - g(x^*) \geq g(x_t) > \varepsilon,$$

we get an estimate for the sum of the objective functionals:

$$\begin{aligned}
\sum_{t \in I} \left(f_t(x_t) - f_t(x^*) \right) &\leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \sum_{t \in J} \left(g(x_t) - g(x^*) \right) \\
&\leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \sum_{t \in J} \varepsilon = \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \varepsilon T_J.
\end{aligned}$$

4. Thus, we get

$$\begin{aligned}
0 \leq \text{Regret}_T &= \sum_{t=1}^T \left(f_t(x_t) - f_t(x^*) \right) = \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \\
&\leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \varepsilon T_J.
\end{aligned}$$

Corollary 1. Assume that all conditions of Theorem 4 hold and suppose $\mu_t \geq \mu > 0$ for all $1 \leq t \leq T + T_J$. If

$$\varepsilon = \frac{M^2}{\mu} \frac{1 + \ln T}{T},$$

then the bound on the regret of Algorithm 4 is $O(\ln T)$.

Proof.

$$0 \leq \text{Regret}_T \leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t}} - \varepsilon T_J \leq \sum_{t=1}^{T+T_J} \frac{M^2}{\mu t} - \varepsilon T_J \leq \frac{M^2}{\mu} \left(\ln(T+T_J) + 1 \right) - \varepsilon T_J,$$

hence $\varepsilon T_J \leq \frac{M^2}{\mu} \left(1 + \ln(T+T_J) \right)$. Let $\varepsilon = \frac{M^2}{\mu} \frac{1 + \ln T}{T}$. Then we have

$$\frac{1 + \ln T}{T} T_J \leq 1 + \ln(T+T_J),$$

and

$$\frac{T_J}{T} \leq \frac{1 + \ln(T+T_J)}{1 + \ln T} = \frac{1 + \ln T + \ln(1 + \frac{T_J}{T})}{1 + \ln T} \leq 1 + \ln(1 + \frac{T_J}{T}).$$

Since the linear function grows faster than the logarithmic one, it is obviously, that with a sufficiently large T_J , the above inequality does not hold, and then $\frac{T_J}{T}$ is bounded. Thus we proved that there exists such a constant $C > 0$, that $\frac{T_J}{T} \leq CT$. So, we have

$$\text{Regret}_T \leq \frac{M^2}{\mu} \left(1 + \ln \left((C+1)T \right) \right) = O(\ln T) = O(T\varepsilon).$$

Let's consider an analogue of Algorithm 4 for relatively strongly convex and relatively Lipschitz-continuous problems with iterative regularization. Similarly to Section 2, we define $\lambda_{1:t} := \sum_{s=1}^t \lambda_s$.

Theorem 5. *Suppose that, for each t , f_t is an M_t -relatively Lipschitz-continuous and μ_t -relatively strongly convex function with respect to the prox-function h . Let $g(x)$ be M_g -relatively Lipschitz-continuous and μ_g -relatively strongly convex function with respect to h . Let $d : Q \rightarrow \mathbb{R}$ be M_d -relatively Lipschitz-continuous and 1-relatively strongly convex function with respect to h . Suppose also that $d(x) \geq 0, \forall x \in Q$. If Algorithm 5 works exactly T productive steps and $\text{Regret}_T \geq 0$, then the following inequalities hold:*

$$\text{Regret}_T \leq \lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J,$$

and

$$\text{Regret}_T \leq 2 \inf_{\lambda_1^*, \dots, \lambda_{T+T_J}^*} \left((A^2 + 2M_d^2) \lambda_{1:T+T_J}^* + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t^* M_d)^2}{\mu_{1:t} + \lambda_{1:t}^*} \right) - \varepsilon T_J.$$

where $A^2 = \sup_{x \in Q} d(x)$, $M = \max\{M_t, M_g\}$ and $g(x_t) \leq \varepsilon$ for any $t = \overline{1}, \overline{T}$.

Algorithm 5 Constrained Online Optimization: Mirror Descent for Relatively Strongly Convex and Relatively Lipschitz-Continuous Problems with Regularization.

Require: $\varepsilon > 0, x_1 \in Q$.

- 1: $i := 1, t := 1$;
- 2: **repeat**
- 3: **if** $g(x_t) \leq \varepsilon$ **then**
- 4: $\lambda_t = \frac{1}{2} \left(\sqrt{(\mu_{1:t} + \lambda_{1:t-1})^2 + 8M^2/(A^2 + 2M_d^2)} - (\mu_{1:t} + \lambda_{1:t-1}) \right)$;
- 5: $\eta_t = \frac{1}{\mu_{1:t} + \lambda_{1:t}}$;
- 6: $x_{t+1} := \text{Pr}_Q \{x_t - \eta_t(\nabla f_t(x_t) + \lambda_t \nabla d(x_t))\}$; "productive step"
- 7: $i := i + 1$;
- 8: $t := t + 1$;
- 9: **else**
- 10: $\lambda_t = \frac{1}{2} \left(\sqrt{(\mu_{1:t} + \lambda_{1:t-1})^2 + 8M^2/(A^2 + 2M_d^2)} - (\mu_{1:t} + \lambda_{1:t-1}) \right)$;
- 11: $\eta_t = \frac{1}{\mu_{1:t} + \lambda_{1:t}}$;
- 12: $x_{t+1} := \text{Pr}_Q \{x_t - \eta_t(\nabla g(x_t) + \lambda_t \nabla d(x_t))\}$; "non-productive step"
- 13: $t := t + 1$;
- 14: **end if**
- 15: **until** $i = T + 1$.
- 16: Guaranteed accuracy:

$$\delta := \frac{1}{T} \left(\lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \right).$$

Proof. The proof is given in Appendix C.

We can formulate the following statement for concrete values of μ_t . Partially, we can achieve intermediate rates for regret between T and $\log T$.

Corollary 2. *Assume that all conditions of Theorem 5 hold and $\mu_t = t^{-\alpha}$ for all $1 \leq t \leq T + T_J$.*

1. *If $\alpha = 0, \lambda_t = 0 \forall 1 \leq t \leq T + T_J$, and $\varepsilon = M^2 \frac{1 + \ln T}{T}$, then the bound on the regret of Algorithm 5 is $O(\ln T)$.*
2. *If $\alpha > 1/2, \lambda_1 = \sqrt{T + T_J}, \lambda_t = 0$ for $1 < t \leq T + T_J$, and*

$$\varepsilon = \frac{A^2 + 2(M_d^2 + M^2)}{\sqrt{T}},$$

then the bound on the regret of Algorithm 5 is $O(\sqrt{T})$.

3. *If $0 < \alpha \leq 1/2, \lambda_1 = (T + T_J)^\alpha, \lambda_t = 0 \forall 1 \leq t \leq T + T_J$ and*

$$\varepsilon = \left(A^2 + 2M_d^2 + \frac{4M^2}{\alpha} \right) T^{\alpha-1},$$

then the bound on the regret of Algorithm 5 is $O(T^\alpha)$.

Proof. The proof is given in Appendix D.

4 Numerical Experiments

In this section, to demonstrate the performance of the proposed Algorithm 4, we conduct some numerical experiments for the considered problem (1) and make a comparison with an adaptive Algorithm 2, proposed in [16]. 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). RAM of the computer is 8 GB.

Let us consider the following function

$$f(x) = \frac{1}{T} \sum_{i=1}^T \left(|\langle a_i, x \rangle - b_i| + \frac{\mu_i}{2} \|x\|_2^2 \right), \quad (14)$$

where $a_i \in \mathbb{R}^n, b_i \in \mathbb{R}, \mu_i > 0$. Functional constraints are defined as follows

$$g(x) = \max_{1 \leq i \leq m} \left\{ \langle \alpha_i, x \rangle - \beta_i + \frac{\hat{\mu}_i}{2} \|x\|_2^2 \right\}, \quad (15)$$

where $\alpha_i \in \mathbb{R}^n, \beta_i \in \mathbb{R}, \hat{\mu}_i > 0$.

Function f is the arithmetic mean of the functions $f_i(x) = |\langle a_i, x \rangle - b_i| + \frac{\mu_i}{2} \|x\|_2^2, i = \overline{1, T}$. Each of these functions is M_i -Lipschitz-continuous and μ_i -strongly convex. Also, function g is M_g -Lipschitz-continuous and μ_g -strongly convex. Coefficients $a_i, \alpha_i \in \mathbb{R}^n$ and constants $b_i, \beta_i \in \mathbb{R}$ in (14) and (15) are randomly generated from the uniform distribution over $[0, 1)$. Also, the strong convexity parameters μ_i and $\hat{\mu}_i$ are randomly chosen in the interval $(0, 1)$.

We choose a standard Euclidean proximal setup as a prox-function, starting point $x_0 = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right) \in \mathbb{R}^n$ and Q is the unit ball in \mathbb{R}^n .

We run Algorithm 4 and adaptive Algorithm 2 from [16] with $n = 1000$ and $m = 10$ and different values of T with $\varepsilon = 1/\sqrt{T}$. The results of the work of these algorithms are represented in Fig. 1, below. These results demonstrate the number of non-productive steps, the running time is given in seconds, the guaranteed accuracy δ of the approximated solution (sequence $\{x_t\}_{t \in I}$ on productive steps), and the values $\frac{1}{T} \sum_{i=1}^T f_i(x_i)$, where x_i is productive, as a function of T . The dotted curve represents the results of the proposed Algorithm 4, whereas the dashed curve represents the results of the adaptive Algorithm 2 in [16].

From the conducted experiments, we can see that the adaptive Algorithm 2 in [16], works faster than Algorithm 4, with a smaller amount of non-productive steps. But when increasing the number of functionals f_i in (14), the guaranteed accuracy δ and values of the objective function at productive steps, produced by Algorithm 4 is better.

Note that from Fig. 1, we can see that increasing of T (the number of functionals f_i) leads to an increasing of δ (the accuracy of the solution). In other words, increasing the number of functionals f_i in the objective function (14), which in fact is increasing information about the objective function or actually enlarging data about the problem, leads to increasing the accuracy of the solution.

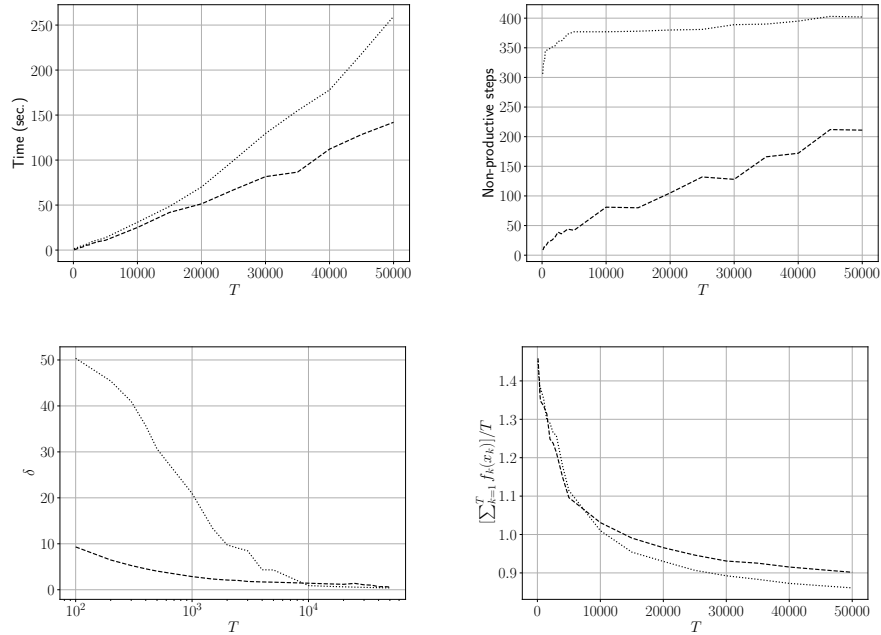


Fig. 1: The results of Algorithm 4 (dots) and adaptive Algorithm 2 in [16] (dashed) for the objective function (14) with constraints (15).

Conclusions

In this paper, we considered relatively strongly convex and relatively Lipschitz-continuous constrained online optimization problems. We proposed some methods with switching over productive and non-productive steps and provided corresponding estimates of the quality of the solution. We also presented analogues of the methods proposed earlier in [6], for solving relatively strongly convex and relatively Lipschitz-continuous online optimization problems with and without regularization. Furthermore, for the problems with functional constraints, we have proposed a scheme with switching over productive and non-productive steps with adaptive regularization. We also proved that if the algorithm runs exactly T productive steps and has a non-negative regret, then the number of non-productive steps satisfies $T_J \leq CT$, where C is a constant. In particular, for the proposed methods, we obtained some bounds on the algorithm's regret in terms of the number of productive steps made by the algorithm under specific assumptions about the parameters of relative strong convexity and some other parameters of the problem.

The key idea of the considered methods is that at each step of the algorithm for each selected f_t , we determine the corresponding parameter of the relative strong convexity μ_t . Thus, it is possible to take into account the parameter of

relative strong convexity of each of the functions f_t . This is highly significant because the functions are selected during the method's working process, and it would be a mistake to assume that some strong convexity can be set initially. It is important to note, that if we consider the following functional constraint

$$g(x) = \max_{1 \leq i \leq m} \{g_i(x)\},$$

where each g_i is μ_i -relatively strongly convex function, then in the process of working of the algorithm at this particular non-productive step t , it makes sense to consider the first of the constraints $g_i(x)$ for which the condition $g_i(x_t) \leq \varepsilon$ is violated and the corresponding parameter μ_i , i.e. $\mu_t = \mu_i$. We do not initially know which constraint will be violated in the process of working of the method, and it is logical to take into account its relative strong convexity parameter instead of the global relative strong convexity one, which may turn out to be much larger. We have analyzed the results of the given numerical experiments and compared the effectiveness of one of the proposed algorithms with Algorithm 2 proposed in [16].

References

1. Alkousa, M. S.: On Some Stochastic Mirror Descent Methods for Constrained Online Optimization Problems. *Computer Research and Modeling*, **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. *Mathematics of Operations Research*, **42**(2), 330–348 (2017)
3. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundation and Trends in Machine Learning*, **5**(1), 1–122 (2012)
4. Gasnikov, A. V., Lagunovskaya, A. A., Usmanova, I. N., Fedorenko, F. A., Krymova, E. A.: Stochastic online optimization. Single-point and multi-point nonlinear multi-armed bandits. Convex and strongly-convex case. *Automation and Remote Control*, **78**(2), 224–234 (2017)
5. Lu, H.: Relative Continuity for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. *Inf. Jour. Opt.*, **1**(4), 288–303 (2019)
6. Hazan, E., Rakhlin, A., Bartlett, P.: Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, **20**, (2007)
7. Hazan, E., Kale, S.: Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *JMLR*. **15** 2489–2512 (2014)
8. Hazan, E: Introduction to online convex optimization. *Foundations and Trends in Optimization*, **2**(3–4)2, 157–325 (2015)
9. Lu, H.: Relative Continuity for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. *Inf. Jour. on Optimization* **1**(4), 288–303 (2019)
10. Lu, H., Freund, R., Nesterov, Yu.: Relatively smooth convex optimization by first-order methods and applications. *SIOPT* **28**(1), 333–354 (2018)

11. Lugosi, G., Cesa-Bianchi, N.: Prediction, learning and games. New York, Cambridge University Press, (2006).
12. Nesterov, Yu.: Relative Smoothness: New Paradigm in Convex Optimization. Conference report, EUSIPCO-2019, A Coruna, Spain, September 4, 2019. <http://eusipco2019.org/wp-content/uploads/2019/10/Relative-Smoothness-New-Paradigm-in-Convex.pdf>
13. Polyak, B. T.: Introduction to optimization. Optimization Software, Inc, New York (1987)
14. Orabona, F., Crammer, K., Cesa-Bianchi, N.: A generalized online mirror descent with applications to classification and regression. *Mach Learn* 99, 411—435 (2015)
15. Stonyakin, F., Titov, A., Alkousa, M., Savchuk, O., Gasnikov, A.: Adaptive Algorithms for Relatively Lipschitz Continuous Convex Optimization Problems. arXiv preprint, <https://arxiv.org/abs/2107.05765> (2021)
16. Titov, A. A., Stonyakin, F. S., Gasnikov, A. V., Alkousa, M. S.: Mirror descent and constrained online optimization problems. In *Optimization and Applications: 9th International Conference, OPTIMA 2018, Petrovac, Montenegro, October 1–5, 2018*, Springer International Publishing. Revised Selected Papers 9, 64–78, (2019)
17. Titov, A. A., Stonyakin, F. S., Alkousa, M. S., Ablav, S. S., Gasnikov, A. V.: Analogues of switching subgradient schemes for relatively Lipschitz-continuous convex programming problems. In *Mathematical Optimization Theory and Operations Research: 19th International Conference, MOTOR 2020, Novosibirsk, Russia, July 6–10, 2020*, Cham: Springer International Publishing. Revised Selected Papers 133–149, (2020)

Appendix A. The proof of Theorem 2.

Proof. By the assumption on the functions f_t , for $x^* = \arg \min_{x \in Q} \sum_{t=1}^T f_t(x)$ we have

$$f_t(x_t) - f_t(x^*) \leq \langle \nabla f_t(x_t), x_t - x^* \rangle - \mu_t V(x^*, x_t).$$

By a well-known property of Bregman divergences, it holds that for any vectors x, y, z ,

$$\langle x - y, \nabla h(z) - \nabla h(y) \rangle = V(x, y) - V(x, z) + V(y, z).$$

Combining both observations,

$$\begin{aligned} f_t(x_t) - f_t(x^*) &\leq \langle \nabla f_t(x_t), x_t - x^* \rangle - \mu_t V(x^*, x_t) \\ &= \frac{1}{\eta_{t+1}} \langle \nabla h(y_{t+1}) - \nabla h(x_t), x^* - x_t \rangle - \mu_t V(x^*, x_t) \\ &= \frac{1}{\eta_{t+1}} [V(x^*, x_t) - V(x^*, y_{t+1}) + V(x_t, y_{t+1})] - \mu_t V(x^*, x_t) \\ &\leq \frac{1}{\eta_{t+1}} [V(x^*, x_t) - V(x^*, x_{t+1}) + V(x_t, y_{t+1})] - \mu_t V(x^*, x_t), \end{aligned}$$

where the last inequality follows from the Pythagorean Theorem for Bregman divergences, as x_{t+1} is the projection w.r.t the Bregman divergence of y_{t+1} and $x^* \in Q$ is in the convex set.

Summing over all iterations and recalling that $\eta_{t+1} = \frac{1}{\mu_{1:t}}$,

$$\begin{aligned} \text{Regret}_T &\leq \sum_{t=2}^T V(x^*, x_t) \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \mu_t \right) + V(x^*, x_1) \left(\frac{1}{\eta_2} - \mu_1 \right) \\ &\quad + \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}) = \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}). \end{aligned} \tag{16}$$

We proceed to bound $V(x_t, y_{t+1})$. By the definition of Bregman divergence, and the M_t -relative Lipschitz-continuity,

$$\begin{aligned} V(x_t, y_{t+1}) + V(y_{t+1}, x_t) &= \langle \nabla h(x_t) - \nabla h(y_{t+1}), x_t - y_{t+1} \rangle \\ &= \eta_{t+1} \langle \nabla f_t(x_t), x_t - y_{t+1} \rangle \\ &\leq \eta_{t+1} M_t \sqrt{2V(y_{t+1}, x_t)} \\ &= \sqrt{2M_t^2 \eta_{t+1}^2 V(y_{t+1}, x_t)} \\ &\leq M_t^2 \eta_{t+1}^2 + V(y_{t+1}, x_t). \end{aligned}$$

Thus, we have

$$V(x_t, y_{t+1}) \leq M_t^2 \eta_{t+1}^2.$$

Plugging back into (16) we get

$$\text{Regret}_T \leq \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}) \leq \sum_{t=1}^T \eta_{t+1} \cdot M_t^2 = \sum_{t=1}^T \frac{M_t^2}{\mu_{1:t}}.$$

Appendix B. The proof of Theorem 3.

At the first, let us mention the following auxiliary lemma, which was proposed in [6].

Lemma 1. *Define*

$$H_T(\{\lambda_t\}) = H_T(\lambda_1, \dots, \lambda_T) = \lambda_{1:T} + \sum_{t=1}^T \frac{C_t}{\mu_{1:t} + \lambda_{1:t}},$$

where $C_t \geq 0$ does not depend on λ_t . If λ_t satisfies $\lambda_t = \frac{C_t}{\mu_{1:t} + \lambda_{1:t}}$ for $t = 1, \dots, T$, then

$$H_T(\{\lambda_t\}) \leq 2 \inf_{\{\lambda_t^*\} \geq 0} H_T(\{\lambda_t^*\}).$$

Now, let us prove Theorem 3.

Proof. By assumption on the functions f_t and d , for $x^* = \arg \min_{x \in Q} \sum_{t=1}^T f_t(x)$ we have

$$f_t(x_t) - f_t(x^*) \leq \langle \nabla f_t(x_t), x_t - x^* \rangle - \mu_t V(x^*, x_t),$$

and

$$d(x_t) - d(x^*) \leq \langle \nabla d(x_t), x_t - x^* \rangle - V(x^*, x_t).$$

Summing these two inequalities, we have

$$\begin{aligned} (f_t(x_t) + \lambda_t d(x_t)) - (f_t(x^*) + \lambda_t d(x^*)) &\leq \langle \nabla f_t(x_t) + \lambda_t \nabla d(x_t), x_t - x^* \rangle \\ &\quad - (\mu_t + \lambda_t) V(x^*, x_t). \end{aligned}$$

By a well-known property of Bregman divergences, it holds that for any vectors x, y, z ,

$$\langle x - y, \nabla h(z) - \nabla h(y) \rangle = V(x, y) - V(x, z) + V(y, z).$$

Combining both observations,

$$\begin{aligned} &(f_t(x_t) + \lambda_t d(x_t)) - (f_t(x^*) + \lambda_t d(x^*)) \\ &\leq \langle \nabla f_t(x_t) + \lambda_t \nabla d(x_t), x_t - x^* \rangle - (\mu_t + \lambda_t) V(x^*, x_t) \\ &= \frac{1}{\eta_{t+1}} \langle \nabla h(y_{t+1}) - \nabla h(x_t), x^* - x_t \rangle - (\mu_t + \lambda_t) V(x^*, x_t) \\ &= \frac{1}{\eta_{t+1}} [V(x^*, x_t) - V(x^*, y_{t+1}) + V(x_t, y_{t+1})] - (\mu_t + \lambda_t) V(x^*, x_t) \\ &\leq \frac{1}{\eta_{t+1}} [V(x^*, x_t) - V(x^*, x_{t+1}) + V(x_t, y_{t+1})] - (\mu_t + \lambda_t) V(x^*, x_t), \end{aligned}$$

where the last inequality follows from the Pythagorean theorem for Bregman divergences, as x_{t+1} is the projection w.r.t the Bregman divergence of y_{t+1} and $x^* \in Q$ is in the convex set.

Summing over all iterations and recalling that $\eta_{t+1} = \frac{1}{\mu_{1:t} + \lambda_{1:t}}$,

$$\begin{aligned}
 & \sum_{t=1}^T (f_t(x_t) + \lambda_t d(x_t)) - \sum_{t=1}^T (f_t(x^*) + \lambda_t d(x^*)) \\
 & \leq \sum_{t=2}^T V(x^*, x_t) \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \mu_t - \lambda_t \right) + V(x^*, x_1) \left(\frac{1}{\eta_2} - \mu_1 - \lambda_1 \right) \quad (17) \\
 & \quad + \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}) = \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}).
 \end{aligned}$$

We proceed to bound $V(x_t, y_{t+1})$. By the definition of Bregman divergence, and the relative Lipschitz-continuity,

$$\begin{aligned}
 V(x_t, y_{t+1}) + V(y_{t+1}, x_t) &= \langle \nabla h(x_t) - \nabla h(y_{t+1}), x_t - y_{t+1} \rangle \\
 &= \eta_{t+1} \langle \nabla f_t(x_t) + \lambda_t \nabla d(x_t), x_t - y_{t+1} \rangle \\
 &\leq \eta_{t+1} M_t \sqrt{2V(y_{t+1}, x_t)} + \lambda_t \eta_{t+1} M_d \sqrt{2V(y_{t+1}, x_t)} \\
 &= (M_t + \lambda_t M_d) \sqrt{2\eta_{t+1}^2 V(y_{t+1}, x_t)} \\
 &= \sqrt{2(M_t + \lambda_t M_d)^2 \eta_{t+1}^2 V(y_{t+1}, x_t)} \\
 &\leq (M_t + \lambda_t M_d)^2 \eta_{t+1}^2 + V(y_{t+1}, x_t).
 \end{aligned}$$

Thus, we have

$$V(x_t, y_{t+1}) \leq (M_t + \lambda_t M_d)^2 \eta_{t+1}^2.$$

Plugging back into (17) we get

$$\begin{aligned}
 & \sum_{t=1}^T (f_t(x_t) + \lambda_t d(x_t)) - \sum_{t=1}^T (f_t(x^*) + \lambda_t d(x^*)) \leq \sum_{t=1}^T \frac{1}{\eta_{t+1}} V(x_t, y_{t+1}) \leq \\
 & \leq \sum_{t=1}^T \eta_{t+1} (M_t + \lambda_t M_d)^2 = \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}}.
 \end{aligned}$$

Thus, we have

$$\sum_{t=1}^T (f_t(x_t) + \lambda_t d(x_t)) \leq \min_x \left(\sum_{t=1}^T (f_t(x) + \lambda_t d(x)) \right) + \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}}.$$

Dropping the $d(x_t)$ terms and bounding $d(x^*) \leq A^2$, we have

$$\sum_{t=1}^T f_t(x_t) \leq \sum_{t=1}^T f_t(x^*) + \lambda_{1:T} A^2 + \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}},$$

hence

$$\text{Regret}_T \leq \lambda_{1:T} A^2 + \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}}. \quad (18)$$

The following inequality allows us to remove the dependence on λ_t from the numerator of the second sum in (18). We have

$$\begin{aligned} \lambda_{1:T}A^2 + \sum_{t=1}^T \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} &\leq \lambda_{1:T}A^2 + \sum_{t=1}^T \left(\frac{2M_t^2}{\mu_{1:t} + \lambda_{1:t}} + \frac{2\lambda_t^2 M_d^2}{\mu_{1:t} + \lambda_{1:t-1} + \lambda_t} \right) \\ &\leq (A^2 + 2M_d^2)\lambda_{1:T} + 2 \sum_{t=1}^T \frac{M_t^2}{\mu_{1:t} + \lambda_{1:t}}. \end{aligned} \quad (19)$$

By (19) and Lemma 1, we have

$$\begin{aligned} \text{Regret}_T &\leq (A^2 + 2M_d^2)\lambda_{1:T} + 2 \sum_{t=1}^T \frac{M_t^2}{\mu_{1:t} + \lambda_{1:t}} \\ &\leq \inf_{\lambda_1^*, \dots, \lambda_T^*} \left(2(A^2 + 2M_d^2)\lambda_{1:T}^* + 4 \sum_{t=1}^T \frac{M_t^2}{\mu_{1:t} + \lambda_{1:t}^*} \right) \\ &\leq 2 \inf_{\lambda_1^*, \dots, \lambda_T^*} \left((A^2 + 2M_d^2)\lambda_{1:T}^* + \sum_{t=1}^T \frac{(M_t + \lambda_t^* M_d)^2}{\mu_{1:t} + \lambda_{1:t}^*} \right), \end{aligned}$$

provided the λ_t are chosen as solutions to

$$(A^2 + 2M_d^2)\lambda_t = \frac{2M_t^2}{\mu_{1:t} + \lambda_{1:t-1} + \lambda_t}.$$

It is easy to verify that

$$\lambda_t = \frac{1}{2} \left(\sqrt{(\mu_{1:t} + \lambda_{1:t-1})^2 + 8M_t^2/(A^2 + 2M_d^2)} - (\mu_{1:t} + \lambda_{1:t-1}) \right)$$

is the non-negative root of the above quadratic equation.

Appendix C. The proof of Theorem 5.

Proof. By assumption on the functions f_t and d for every productive step we have

$$\begin{aligned} &\eta_t((f_t(x_t) + \lambda_t d(x_t)) - (f_t(x^*) + \lambda_t d(x^*))) \\ &\leq \eta_t(\langle \nabla f_t(x_t) + \lambda_t \nabla d(x_t), x_t - x^* \rangle - (\mu_t + \lambda_t)V(x^*, x_t)) \\ &\leq \eta_t^2(M_t + \lambda_t M_d)^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t(\mu_t + \lambda_t)V(x^*, x_t). \end{aligned}$$

Hence, after dividing both sides of the above inequality by η_t we get

$$\begin{aligned}
 & (f_t(x_t) + \lambda_t d(x_t)) - (f_t(x^*) + \lambda_t d(x^*)) \\
 & \leq \eta_t (M_t + \lambda_t M_d)^2 + \frac{1}{\eta_t} (V(x^*, x_t) - V(x^*, x_{t+1})) - (\mu_t + \lambda_t) V(x^*, x_t) \\
 & = \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_t) - (\mu_t + \lambda_t) V(x^*, x_t) - \\
 & \quad - (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_{t+1}) \\
 & = \frac{(M_t + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + (\mu_{1:t-1} + \lambda_{1:t-1}) V(x^*, x_t) - (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_{t+1}).
 \end{aligned}$$

Similarly, taking into account the M_g -relative Lipschitz-continuity of g and the M_d -relative Lipschitz-continuity of d for every non-productive step we have $g(x_t) > \varepsilon$, and

$$\begin{aligned}
 \eta_t \varepsilon & < \eta_t ((g(x_t) + \lambda_t d(x_t)) - (g(x^*) + \lambda_t d(x^*))) \\
 & \leq \eta_t ((\nabla g(x_t) + \lambda_t \nabla d(x_t), x_t - x^*) - (\mu_t + \lambda_t) V(x^*, x_t)) \\
 & \leq \eta_t^2 (M_g + \lambda_t M_d)^2 + V(x^*, x_t) - V(x^*, x_{t+1}) - \eta_t (\mu_t + \lambda_t) V(x^*, x_t).
 \end{aligned}$$

Dividing both sides of the last inequality by η_t , we get:

$$\begin{aligned}
 \varepsilon & < (g(x_t) + \lambda_t d(x_t)) - (g(x^*) + \lambda_t d(x^*)) \\
 & \leq \eta_t (M_g + \lambda_t M_d)^2 + \frac{1}{\eta_t} (V(x^*, x_t) - V(x^*, x_{t+1})) - (\mu_t + \lambda_t) V(x^*, x_t) \\
 & = \frac{(M_g + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_t) - (\mu_t + \lambda_t) V(x^*, x_t) - \\
 & \quad - (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_{t+1}) \\
 & = \frac{(M_g + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + (\mu_{1:t-1} + \lambda_{1:t-1}) V(x^*, x_t) - (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_{t+1}).
 \end{aligned}$$

Summing up the inequalities for productive and non-productive steps, and let $M = \max\{M_t, M_g\}$, then

$$\begin{aligned}
 & \sum_{t \in I} ((f_t(x_t) + \lambda_t d(x_t)) - (f_t(x^*) + \lambda_t d(x^*))) + \sum_{t \in J} ((g(x_t) + \lambda_t d(x_t)) - (g(x^*) + \lambda_t d(x^*))) \\
 & \leq \sum_{t=1}^{T+T_J} \left(\frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + (\mu_{1:t-1} + \lambda_{1:t-1}) V(x^*, x_t) - (\mu_{1:t} + \lambda_{1:t}) V(x^*, x_{t+1}) \right) \\
 & \leq \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - (\mu_{1:T+T_J} + \lambda_{1:T+T_J}) V(x^*, x_{T+T_J}) \\
 & \leq \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}}.
 \end{aligned}$$

Bounding $d(x^*) \leq A^2$ and using the fact, that for non-productive steps

$$g(x_t) - g(x^*) \geq g(x_t) > \varepsilon,$$

we get an estimate for the sum of the objective functionals:

$$\begin{aligned} \sum_{t=1}^T (f_t(x_t) - f_t(x^*)) &= \sum_{t=1}^T f_t(x_t) - \min_{x \in Q} \sum_{t=1}^T f_t(x) \\ &\leq \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \sum_{t \in J} (g(x_t) - g(x^*)) \\ &\quad + \sum_{t=1}^{T+T_J} \lambda_t d(x^*) - \sum_{t=1}^{T+T_J} \lambda_t d(x_t) \\ &\leq \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + \lambda_{1:T+T_J} A^2 - \varepsilon T_J. \end{aligned}$$

Thus, we get

$$0 \leq \text{Regret}_T \leq \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} + \lambda_{1:T+T_J} A^2 - \varepsilon T_J.$$

Using inequality (19), we have

$$\begin{aligned} &\lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\ &\leq \lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \left(\frac{2M^2}{\mu_{1:t} + \lambda_{1:t}} + \frac{2\lambda_t^2 M_d^2}{\mu_{1:t} + \lambda_{1:t-1} + \lambda_t} \right) - \varepsilon T_J \quad (20) \\ &\leq (A^2 + 2M_d^2) \lambda_{1:T+T_J} + 2 \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J. \end{aligned}$$

By (20) and Lemma 1

$$\begin{aligned} \text{Regret}_T &\leq (A^2 + 2M_d^2) \lambda_{1:T+T_J} + 2 \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\ &\leq \inf_{\lambda_1^*, \dots, \lambda_{T+T_J}^*} \left(2(A^2 + 2M_d^2) \lambda_{1:T+T_J}^* + 4 \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t} + \lambda_{1:t}^*} \right) - \varepsilon T_J \\ &\leq 2 \inf_{\lambda_1^*, \dots, \lambda_{T+T_J}^*} \left((A^2 + 2M_d^2) \lambda_{1:T+T_J}^* + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t^* M_d)^2}{\mu_{1:t} + \lambda_{1:t}^*} \right) - \varepsilon T_J. \end{aligned}$$

provided the λ_t are chosen as solutions to

$$(A^2 + 2M_d^2) \lambda_t = \frac{2M^2}{\mu_{1:t} + \lambda_{1:t-1} + \lambda_t}.$$

It is easy to verify that

$$\lambda_t = \frac{1}{2} \left(\sqrt{(\mu_{1:t} + \lambda_{1:t-1})^2 + 8M^2/(A^2 + 2M_d^2)} - (\mu_{1:t} + \lambda_{1:t-1}) \right)$$

is the non-negative root of the above quadratic equation.

Appendix D. The proof of Corollary 2.

Proof. 1. Indeed, if $\lambda_t = 0 \forall 1 \leq t \leq T + T_J$, then the claimed statement immediately follows from Corollary 1.

2. Indeed, if $\lambda_1 = \sqrt{T + T_J}$ and $\lambda_t = 0$ for $1 < t \leq T + T_J$, then

$$\begin{aligned} 0 \leq \text{Regret}_T &\leq \lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\ &\leq (A^2 + 2M_d^2) \lambda_{1:T+T_J} + 2 \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\ &\leq (A^2 + 2M_d^2) \sqrt{T + T_J} + 2 \sum_{t=1}^{T+T_J} \frac{M^2}{\sqrt{T + T_J}} - \varepsilon T_J \\ &= (A^2 + 2(M_d^2 + M^2)) \sqrt{T + T_J} - \varepsilon T_J, \end{aligned}$$

hence $\varepsilon T_J \leq (A^2 + 2(M_d^2 + M^2)) \sqrt{T + T_J}$. Let $\varepsilon = \frac{A^2 + 2(M_d^2 + M^2)}{\sqrt{T}}$. Then we get

$$\frac{T_J}{T} \leq \sqrt{\frac{T + T_J}{T}} = \sqrt{1 + \frac{T_J}{T}}.$$

Since the linear function grows faster than the square root function, it is obviously, that with a sufficiently large T_J , the above inequality does not hold, and then $\frac{T_J}{T}$ is bounded. Thus we proved that $\exists C > 0 : T_J \leq C \cdot T$. So, we have

$$\text{Regret}_T \leq (A^2 + 2(M_d^2 + M^2)) \sqrt{(C + 1)T} = O(\sqrt{T}).$$

3. Let us assume $\lambda_1 = (T + T_J)^\alpha$, $\lambda_t = 0$, $\forall 1 \leq t \leq T + T_J$. Note that

$$\mu_{1:t} := \sum_{s=1}^t \mu_s \geq \int_0^{t-1} (x+1)^{-\alpha} dx = (1-\alpha)^{-1} (t^{1-\alpha} - 1).$$

Hence

$$\begin{aligned}
0 \leq \text{Regret}_T &\leq \lambda_{1:T+T_J} A^2 + \sum_{t=1}^{T+T_J} \frac{(M + \lambda_t M_d)^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\
&\leq (A^2 + 2M_d^2) \lambda_{1:T+T_J} + 2 \sum_{t=1}^{T+T_J} \frac{M^2}{\mu_{1:t} + \lambda_{1:t}} - \varepsilon T_J \\
&\leq (A^2 + 2M_d^2)(T + T_J)^\alpha + 2M^2(1 - \alpha) \sum_{t=1}^{T+T_J} \frac{1}{(t^{1-\alpha} - 1)} - \varepsilon T_J \\
&\leq (A^2 + 2M_d^2)(T + T_J)^\alpha + 4M^2 \frac{1}{\alpha} (T + T_J)^\alpha + O(1) - \varepsilon T_J.
\end{aligned}$$

Then we have $\varepsilon T_J \leq (A^2 + 2M_d^2 + 4M^2 \frac{1}{\alpha})(T + T_J)^\alpha$. Let $\varepsilon = (A^2 + 2M_d^2 + 4M^2 \frac{1}{\alpha}) \frac{T^\alpha}{T}$, then

$$\frac{T^\alpha}{T} T_J \leq (T + T_J)^\alpha,$$

and

$$\frac{T_J}{T} \leq \left(\frac{T + T_J}{T} \right)^\alpha = \left(1 + \frac{T_J}{T} \right)^\alpha.$$

It is obviously, that with a sufficiently large T_J , the above inequality does not hold, and then $\exists C > 0 : T_J \leq C \cdot T$. Thus, we have

$$\text{Regret}_T = O((T + T_J)^\alpha) = O(((C + 1)T)^\alpha) = O(T^\alpha).$$