Chapter 8
Mirror Descent and Convex Optimization Problems with Non-smooth Inequality Constraints

Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov

Abstract We consider the problem of minimization of a convex function on a simple set with convex non-smooth inequality constraint and describe first-order methods to solve such problems in different situations: smooth or non-smooth objective function; convex or strongly convex objective and constraint; deterministic or randomized information about the objective and constraint. Described methods are based on Mirror Descent algorithm and switching subgradient scheme. One of our focus is to propose, for the listed different settings, a Mirror Descent with adaptive stepsizes and adaptive stopping rule. We also construct Mirror Descent for problems with objective function, which is not Lipschitz, e.g., is a quadratic function. Besides that, we address the question of recovering the dual solution in the considered problem.

A. Bayandina
Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, Russia

P. Dvurechensky (✉)
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Institute for Information Transmission Problems RAS, Moscow, Russia
e-mail: pavel.dvurechensky@wias-berlin.de

A. Gasnikov
Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia

Institute for Information Transmission Problems RAS, Moscow, Russia
e-mail: gasnikov@yandex.ru

F. Stonyakin
V.I. Vernadsky Crimean Federal University, Simferopol, Russia

Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
e-mail: fedyor@mail.ru

A. Titov
Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
e-mail: a.a.titov@phystech.edu

P. Giselsson, A. Rantzer (eds.), *Large-Scale and Distributed Optimization*, Lecture Notes in Mathematics 2227,
https://doi.org/10.1007/978-3-319-97478-1_8

Keywords Constrained non-smooth convex optimization · Stochastic adaptive mirror descent · Primal-dual methods · Restarts

AMS Subject Classifications 90C25, 90C30, 90C06, 65Y20

8.1 Introduction

We consider the problem of minimization of a convex function on a simple set with convex non-smooth inequality constraint and describe first-order methods to solve such problems in different situations: smooth or non-smooth objective function; convex or strongly convex objective and constraint; deterministic or randomized information about the objective and constraint. The reason for considering first-order methods is potential large (more than 10^5) number of decision variables.

Because of the non-smoothness presented in the problem, we consider subgradient methods. These methods have a long history starting with the method for deterministic unconstrained problems and Euclidean setting in [28] and the generalization for constrained problems in [25], where the idea of steps switching between the direction of subgradient of the objective and the direction of subgradient of the constraint was suggested. Non-Euclidean extension, usually referred to as Mirror Descent, originated in [17, 19] and later analyzed in [5]. An extension for constrained problems was proposed in [19], see also recent version in [6]. Mirror Descent for unconstrained stochastic optimization problems was introduced in [16], see also [12, 15], and extended for stochastic optimization problems with expectation constraints in [14]. To prove faster convergence rate of Mirror Descent for strongly convex objective in unconstrained case, the restart technique [18–20] was used in [12]. An alternative approach for strongly convex stochastic optimization problems with strongly convex expectation constraints is used in [14].

Usually, the stepsize and stopping rule for Mirror Descent requires to know the Lipschitz constant of the objective function and constraint, if any. Adaptive stepsizes, which do not require this information, are considered in [7] for problems without inequality constraints, and in [6] for constrained problems. Nevertheless, the stopping criterion, expressed in the number of steps, still requires knowledge of Lipschitz constants. One of our focus in this chapter is to propose, for constrained problems, a Mirror Descent with adaptive stepsizes and adaptive stopping rule. We also adopt the ideas of [21, 24] to construct Mirror Descent for problems with objective function, which is not Lipschitz, e.g., a quadratic function. Another important issue, we address, is recovering the dual solution of the considered problem, which was considered in different contexts in [1, 4, 23].

Formally speaking, we consider the following convex constrained minimization problem

$$\min\{f(x): \quad x \in X \subset E, \quad g(x) \leq 0\}, \tag{8.1}$$

where X is a convex closed subset of a finite-dimensional real vector space E, $f : X \to \mathbb{R}$, $g : E \to \mathbb{R}$ are convex functions.

We assume g to be a non-smooth Lipschitz-continuous function and the problem (8.1) to be regular. The last means that there exists a point $\bar{x}$ in relative interior of the set X, such that $g(\bar{x}) < 0$.

Note that, despite problem (8.1) contains only one inequality constraint, considered algorithms allow to solve more general problems with a number of constraints given as $\{g_i(x) \leq 0, i = 1, \dots, m\}$. The reason is that these constraints can be aggregated and represented as an equivalent constraint given by $\{g(x) \leq 0\}$, where $g(x) = \max_{i=1,\dots,m} g_i(x)$.

The the rest of the chapter is divided in three parts. In Sect. 8.2, we describe some basic facts about Mirror Descent, namely, we define the notion of proximal setup, the Mirror Descent step, and provide the main lemma about the progress on each iteration of this method. Section 8.3 is devoted to deterministic constrained problems, among which we consider convex non-smooth problems, strongly convex non-smooth problems and convex problems with smooth objective. The last, Sect. 8.4, considers randomized setting with available stochastic subgradients for the objective and constraint and possibility to calculate the constraint function. We consider methods for convex and strongly convex problems and provide complexity guarantees in terms of expectation of the objective residual and constraint infeasibility, as long as in terms of large deviation probability for these two quantities.

Notation Given a subset I of natural numbers, we denote $|I|$ the number of its elements.

8.2 Mirror Descent Basics

We consider algorithms, which are based on Mirror Descent method. Thus, we start with the description of proximal setup and basic properties of Mirror Descent step. Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x \rangle$. Let $\|\cdot\|_E$ be some norm on E, $\|\cdot\|_{E,*}$ be its dual, defined by $\|g\|_{E,*} = \max\limits_{x} \big\{\langle g, x \rangle, \|x\|_E \leq 1\big\}$. We use $\nabla f(x)$ to denote any subgradient of a function f at a point $x \in \operatorname{dom} f$.

We choose a prox-function $d(x)$, which is continuous, convex on X and

1. admits a continuous in $x \in X^0$ selection of subgradients $\nabla d(x)$, where $X^0 \subseteq X$ is the set of all x, where $\nabla d(x)$ exists;
2. $d(x)$ is 1-strongly convex on X with respect to $\|\cdot\|_E$, i.e., for any $x \in X^0$, $y \in X$ $d(y) - d(x) - \langle \nabla d(x), y - x \rangle \geq \frac{1}{2}\|y - x\|_E^2$.

Without loss of generality, we assume that $\min\limits_{x \in X} d(x) = 0$.

We define also the corresponding Bregman divergence $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$, $x \in X$, $z \in X^0$. Standard proximal setups, i.e., Euclidean, entropy, ℓ_1/ℓ_2, simplex, nuclear norm, spectahedron can be found in [8].

Given a vector $x \in X^0$, and a vector $p \in E^*$, the Mirror Descent step is defined as

$$x_+ = \text{Mirr}[x](p) := \arg\min_{u \in X} \big\{ \langle p, u \rangle + V[x](u) \big\} \\ = \arg\min_{u \in X} \big\{ \langle p, u \rangle + d(u) - \langle \nabla d(x), u \rangle \big\}. \tag{8.2}$$

We make the simplicity assumption, which means that $\text{Mirr}[x](p)$ is easily computable. The following lemma [7] describes the main property of the Mirror Descent step. We prove it here for the reader convenience and to make the chapter self-contained.

Lemma 1 *Let f be some convex function over a set X, $h > 0$ be a stepsize, $x \in X^0$. Let the point x_+ be defined by $x_+ = \text{Mirr}[x](h \cdot (\nabla f(x) + \Delta))$, where $\Delta \in E^*$. Then, for any $u \in X$,*

$$h \cdot \big(f(x) - f(u) + \langle \Delta, x - u \rangle \big) \le h \cdot \langle \nabla f(x) + \Delta, x - u \rangle \\ \le \frac{h^2}{2} \|\nabla f(x) + \Delta\|_{E,*}^2 + V[x](u) - V[x_+](u). \tag{8.3}$$

Proof By optimality condition in (8.2), we have that there exists a subgradient $\nabla d(x_+)$, such that, for all $u \in X$,

$$\langle h \cdot (\nabla f(x) + \Delta) + \nabla d(x_+) - \nabla d(x), u - x_+ \rangle \ge 0.$$

Hence, for all $u \in X$,

$$\begin{aligned} &\langle h \cdot (\nabla f(x) + \Delta), x - u \rangle \qquad (8.4)\\ &\le \langle h \cdot (\nabla f(x) + \Delta), x - x_+ \rangle + \langle \nabla d(x_+) - \nabla d(x), u - x_+ \rangle \\ &= \langle h \cdot (\nabla f(x) + \Delta), x - x_+ \rangle + (d(u) - d(x) - \langle \nabla d(x), u - x \rangle) \\ &\quad - (d(u) - d(x_+) - \langle \nabla d(x_+), u - x_+ \rangle) \\ &\quad - (d(x_+) - d(x) - \langle \nabla d(x), x_+ - x \rangle) \\ &\le \langle h \cdot (\nabla f(x) + \Delta), x - x_+ \rangle + V[x](u) - V[x_+](u) - \frac{1}{2}\|x_+ - x\|_E^2 \\ &\le V[x](u) - V[x_+](u) + \frac{h^2}{2} \|(\nabla f(x) + \Delta)\|_{E,*}^2, \end{aligned}$$

where we used the fact that, for any $g \in E^*$,

$$\max_{y \in E} \langle g, y \rangle - \frac{1}{2}\|y\|_E^2 = \frac{1}{2}\|g\|_{E,*}^2.$$

By convexity of f, we obtain the left inequality in (8.3). □

8.3 Deterministic Constrained Problems

In this section, we consider problem (8.1) in two different settings, namely, non-smooth Lipschitz-continuous objective function f and general objective function f, which is not necessarily Lipschitz-continuous, e.g., a quadratic function. In both cases, we assume that g is non-smooth and is Lipschitz-continuous

$$|g(x) - g(y)| \leq M_g \|x - y\|_E, \quad x, y \in X. \tag{8.5}$$

Let x_* be a solution to (8.1). We say that a point $\tilde{x} \in X$ is an *ε-solution* to (8.1) if

$$f(\tilde{x}) - f(x_*) \leq \varepsilon, \quad g(\tilde{x}) \leq \varepsilon. \tag{8.6}$$

The methods we describe are based on the of Polyak's switching subgradient method [25] for constrained convex problems, also analyzed in [21], and Mirror Descent method originated in [19]; see also [7].

8.3.1 Convex Non-smooth Objective Function

In this subsection, we assume that f is a non-smooth Lipschitz-continuous function

$$|f(x) - f(y)| \leq M_f \|x - y\|_E, \quad x, y \in X. \tag{8.7}$$

Let x_* be a solution to (8.1) and assume that we know a constant $\Theta_0 > 0$ such that

$$d(x_*) \leq \Theta_0^2. \tag{8.8}$$

For example, if X is a compact set, one can choose $\Theta_0^2 = \max_{x \in X} d(x)$. We further develop line of research [1, 4], but we should also mention close works [6, 23]. In comparison to known algorithms in the literature, the main advantage of our method for solving (8.1) is that the stopping criterion does not require the knowledge of constants M_f, M_g, and, in this sense, the method is adaptive. Mirror Descent with stepsizes not requiring knowledge of Lipschitz constants can be found, e.g., in [7] for problems without inequality constraints, and, for constrained problems, in

Algorithm 1 Adaptive mirror descent (non-smooth objective)

Input: accuracy $\varepsilon > 0$; Θ_0 s.t. $d(x_*) \le \Theta_0^2$.
$x^0 = \arg\min_{x \in X} d(x)$.
Initialize the set I as empty set.
Set $k = 0$.
repeat
  if $g(x^k) \le \varepsilon$ **then**
    $M_k = \|\nabla f(x^k)\|_{E,*}$,
    $h_k = \frac{\varepsilon}{M_k^2}$
    $x^{k+1} = \mathrm{Mirr}[x^k](h_k \nabla f(x^k))$ ("productive step")
    Add k to I.
  else
    $M_k = \|\nabla g(x^k)\|_{E,*}$
    $h_k = \frac{\varepsilon}{M_k^2}$
    $x^{k+1} = \mathrm{Mirr}[x^k](h_k \nabla g(x^k))$ ("non-productive step")
  end if
  Set $k = k + 1$.
until $\sum\limits_{j=0}^{k-1} \frac{1}{M_j^2} \ge \frac{2\Theta_0^2}{\varepsilon^2}$
Output: $\bar{x}^k := \frac{\sum\limits_{i \in I} h_i x^i}{\sum\limits_{i \in I} h_i}$

[6].The algorithm is similar to the one in [2], but, for the sake of consistency with other parts of the chapter, we use slightly different proof.

Theorem 1 *Assume that inequalities* (8.5) *and* (8.7) *hold and a known constant* $\Theta_0 > 0$ *is such that* $d(x_*) \le \Theta_0^2$. *Then, Algorithm 1 stops after not more than*

$$k = \left\lceil \frac{2\max\{M_f^2, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{8.9}$$

iterations and $\bar{x}^k$ *is an* ε*-solution to* (8.1) *in the sense of* (8.6).

Proof First, let us prove that the inequality in the stopping criterion holds for k defined in (8.9). By (8.5) and (8.7), we have that, for any $i \in \{0, \dots, k-1\}$, $M_i \le \max\{M_f, M_g\}$. Hence, by (8.9), $\sum\limits_{j=0}^{k-1} \frac{1}{M_j^2} \ge \frac{k}{\max\{M_f^2, M_g^2\}} \ge \frac{2\Theta_0^2}{\varepsilon^2}$.

Denote $[k] = \{i \in \{0, \dots, k-1\}\}$, $J = [k] \setminus I$. From Lemma 1 with $\Delta = 0$, we have, for all $i \in I$ and all $u \in X$,

$$h_i \cdot \big(f(x^i) - f(u)\big) \le \frac{h_i^2}{2}\|\nabla f(x^i)\|_{E,*}^2 + V[x^i](u) - V[x^{i+1}](u)$$

and, for all $i \in J$ and all $u \in X$,

$$h_i \cdot \big(g(x^i) - g(u)\big) \le \frac{h_i^2}{2}\|\nabla g(x^i)\|_{E,*}^2 + V[x^i](u) - V[x^{i+1}](u).$$

Summing up these inequalities for i from 0 to $k-1$, using the definition of h_i, $i \in \{0, \dots, k-1\}$, and taking $u = x_*$, we obtain

$$\begin{aligned}
&\sum_{i\in I} h_i\big(f(x^i) - f(x_*)\big) + \sum_{i\in J} h_i\big(g(x^i) - g(x_*)\big) \\
&\quad \le \sum_{i\in I} \frac{h_i^2 M_i^2}{2} + \sum_{i\in J} \frac{h_i^2 M_i^2}{2} + \sum_{i\in[k]} \big(V[x^i](x_*) - V[x^{i+1}](x_*)\big) \\
&\quad \le \frac{\varepsilon}{2}\sum_{i\in[k]} h_i + \Theta_0^2. \qquad (8.10)
\end{aligned}$$

We also used that, by definition of x^0 and (8.8),

$$V[x^0](x_*) = d(x_*) - d(x^0) - \langle \nabla d(x^0), x_* - x^0\rangle \le d(x_*) \le \Theta_0^2.$$

Since, for $i \in J$, $g(x^i) - g(x_*) \ge g(x^i) > \varepsilon$, by convexity of f and the definition of $\bar{x}^k$, we have

$$\begin{aligned}
\left(\sum_{i\in I} h_i\right)\big(f(\bar{x}^k) - f(x_*)\big) &\le \sum_{i\in I} h_i \left(f(x^i) - f(x_*)\right) < \frac{\varepsilon}{2}\sum_{i\in[k]} h_i - \varepsilon\sum_{i\in J} h_i + \Theta_0^2 \\
&= \varepsilon\sum_{i\in I} h_i - \frac{\varepsilon^2}{2}\sum_{i\in[k]}\frac{1}{M_i^2} + \Theta_0^2 \le \varepsilon\sum_{i\in I} h_i, \qquad (8.11)
\end{aligned}$$

where in the last inequality, the stopping criterion is used. As long as the inequality is strict, the case of the empty I is impossible. Thus, the point $\bar{x}^k$ is correctly defined. Dividing both parts of the inequality by $\sum_{i\in I} h_i$, we obtain the left inequality in (8.6).

For $i \in I$, it holds that $g(x^i) \le \varepsilon$. Then, by the definition of $\bar{x}^k$ and the convexity of g,

$$g(\bar{x}^k) \le \left(\sum_{i\in I} h_i\right)^{-1} \sum_{i\in I} h_i g(x^i) \le \varepsilon.$$

□

Let us now show that Algorithm 1 allows to reconstruct an approximate solution to the problem, which is dual to (8.1). We consider a special type of problem (8.1) with g given by

$$g(x) = \max_{i \in \{1,\dots,m\}} \big\{ g_i(x) \big\}. \tag{8.12}$$

Then, the dual problem to (8.1) is

$$\varphi(\lambda) = \min_{x \in X} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right\} \to \max_{\lambda_i \geq 0, i=1,\dots,m} \varphi(\lambda), \tag{8.13}$$

where $\lambda_i \geq 0, i = 1, \dots, m$ are Lagrange multipliers.

We slightly modify the assumption (8.8) and assume that the set X is bounded and that we know a constant $\Theta_0 > 0$ such that

$$\max_{x \in X} d(x) \leq \Theta_0^2.$$

As before, denote $[k] = \{j \in \{0, \dots, k-1\}\}$, $J = [k] \setminus I$. Let $j \in J$. Then a subgradient of $g(x)$ is used to make the j-th step of Algorithm 1. To find this subgradient, it is natural to find an active constraint $i \in 1, \dots, m$ such that $g(x^j) = g_i(x^j)$ and use $\nabla g(x^j) = \nabla g_i(x^j)$ to make a step. Denote $i(j) \in 1, \dots, m$ the number of active constraint, whose subgradient is used to make a non-productive step at iteration $j \in J$. In other words, $g(x^j) = g_{i(j)}(x^j)$ and $\nabla g(x^j) = \nabla g_{i(j)}(x^j)$. We define an approximate dual solution on a step $k \geq 0$ as

$$\bar{\lambda}_i^k = \frac{1}{\sum\limits_{j \in I} h_j} \sum_{j \in J, i(j) = i} h_j, \quad i \in \{1, \dots, m\}. \tag{8.14}$$

and modify Algorithm 1 to return a pair $(\bar{x}^k, \bar{\lambda}^k)$.

Theorem 2 *Assume that the set X is bounded, the inequalities (8.5) and (8.7) hold and a known constant $\Theta_0 > 0$ is such that $d(x_*) \leq \Theta_0^2$. Then, modified Algorithm 1 stops after not more than*

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

iterations and the pair $(\bar{x}^k, \bar{\lambda}^k)$ returned by this algorithm satisfies

$$f(\bar{x}^k) - \varphi(\bar{\lambda}^k) \leq \varepsilon, \quad g(\bar{x}^k) \leq \varepsilon. \tag{8.15}$$

Proof From Lemma 1 with $\Delta = 0$, we have, for all $j \in I$ and all $u \in X$,

$$h_j \big(f(x^j) - f(u) \big) \leq \frac{h_j^2}{2} \|\nabla f(x^j)\|_{E,*}^2 + V[x^j](u) - V[x^{j+1}](u)$$

and, for all $j \in J$ and all $u \in X$,

$$\begin{aligned} h_j\big(g_{i(j)}(x^j) - g_{i(j)}(u)\big) &\le h_j \langle \nabla g_{i(j)}(x^j), x^j - u\rangle \\ &= h_j \langle \nabla g(x^j), x^j - u\rangle \\ &\le \frac{h_j^2}{2}\|\nabla g(x^j)\|_{E,*}^2 + V[x^j](u) - V[x^{j+1}](u). \end{aligned}$$

Summing up these inequalities for j from 0 to $k-1$, using the definition of h_j, $j \in \{0, \dots, k-1\}$, we obtain, for all $u \in X$,

$$\begin{aligned} &\sum_{j\in I} h_j\big(f(x^j) - f(u)\big) + \sum_{j\in J} h_j\big(g_{i(j)}(x^j) - g_{i(j)}(u)\big) \\ &\quad \le \sum_{i\in I} \frac{h_j^2 M_j^2}{2} + \sum_{j\in J} \frac{h_j^2 M_j^2}{2} + \sum_{j\in[k]} \big(V[x^j](u) - V[x^{j+1}](u)\big) \\ &\quad \le \frac{\varepsilon}{2}\sum_{j\in[k]} h_j + \Theta_0^2. \end{aligned}$$

Since, for $j \in J$, $g_{i(j)}(x^j) = g(x^j) > \varepsilon$, by convexity of f and the definition of $\bar{x}^k$, we have, for all $u \in X$,

$$\begin{aligned} \left(\sum_{j\in I} h_j\right)\big(f(\bar{x}^k) - f(u)\big) &\le \sum_{j\in I} h_j \left(f(x^j) - f(u)\right) \\ &\le \frac{\varepsilon}{2}\sum_{j\in[k]} h_j + \Theta_0^2 - \sum_{j\in J} h_j\big(g_{i(j)}(x^j) - g_{i(j)}(u)\big) \\ &< \frac{\varepsilon}{2}\sum_{j\in[k]} h_i + \Theta_0^2 - \varepsilon\sum_{j\in J} h_i + \sum_{j\in J} h_j g_{i(j)}(u) \\ &= \varepsilon\sum_{j\in I} h_j - \frac{\varepsilon^2}{2}\sum_{j\in[k]}\frac{1}{M_j^2} + \Theta_0^2 + \sum_{j\in J} h_j g_{i(j)}(u) \\ &\le \varepsilon\sum_{j\in I} h_j + \sum_{j\in J} h_j g_{i(j)}(u), \end{aligned} \tag{8.16}$$

where in the last inequality, the stopping criterion is used. At the same time, by (8.14), for all $u \in X$,

$$\sum_{j\in J} h_j g_{i(j)}(u) = \sum_{i=1}^{m}\sum_{j\in J, i(j)=i} h_j g_{i(j)}(u) = \left(\sum_{j\in I} h_j\right)\sum_{i=1}^{m} \bar{\lambda}_i^k g_i(u).$$

This and (8.16) give, for all $u \in X$,

$$\left(\sum_{j \in I} h_j\right) f(\bar{x}^k) < \left(\sum_{j \in I} h_j\right) \left(f(u) + \varepsilon + \sum_{i=1}^{m} \bar{\lambda}_i^k g_i(u)\right).$$

Since the inequality is strict and holds for all $u \in X$, we have $\left(\sum_{j \in I} h_j\right) \neq 0$ and

$$\begin{aligned} f(\bar{x}^k) &< \varepsilon + \min_{u \in X} \left\{ f(u) + \sum_{i=1}^{m} \bar{\lambda}_i^k g_i(u) \right\} \\ &= \varepsilon + \varphi(\bar{\lambda}^k). \end{aligned} \tag{8.17}$$

Second inequality in (8.15) follows from Theorem 1. □

8.3.2 Strongly Convex Non-smooth Objective Function

In this subsection, we consider problem (8.1) with assumption (8.7) and additional assumption of strong convexity of f and g with the same parameter μ, i.e.,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_E^2, \quad x, y \in X$$

and the same holds for g. For example, $f(x) = x^2 + |x|$ is a Lipschitz-continuous and strongly convex function on $X = [-1; 1] \subset \mathbb{R}$. We also slightly modify assumptions on prox-function $d(x)$. Namely, we assume that $0 = \arg\min_{x \in X} d(x)$ and that d is bounded on the unit ball in the chosen norm $\| \cdot \|_E$, that is

$$d(x) \leq \frac{\Omega}{2}, \quad \forall x \in X : \|x\|_E \leq 1, \tag{8.18}$$

where Ω is some known number. Finally, we assume that we are given a starting point $x_0 \in X$ and a number $R_0 > 0$ such that $\|x_0 - x_*\|_E^2 \leq R_0^2$.

To construct a method for solving problem (8.1) under stated assumptions, we use the idea of restarting Algorithm 1. The idea of restarting a method for convex problems to obtain faster rate of convergence for strongly convex problems dates back to 1980s, see [19, 20]. The algorithm is similar to the one in [2], but, for the sake of consistency with other parts of the chapter, we use slightly different proof. To show that restarting algorithm is also possible for problems with inequality constraints, we rely on the following lemma.

Lemma 2 *Let f and g be strongly convex functions with the same parameter μ and x_* be a solution of the problem (8.1). If, for some $\tilde{x} \in X$,*

$$f(\tilde{x}) - f(x_*) \le \varepsilon, \quad g(\tilde{x}) \le \varepsilon,$$

then

$$\frac{\mu}{2}\|\tilde{x} - x_*\|_E^2 \le \varepsilon.$$

Proof Since problem (8.1) is regular, by necessary optimality condition [9] at the point x_*, there exist $\lambda_0, \lambda \ge 0$ not equal to 0 simultaneously, and subgradients $\nabla f(x_*)$, $\nabla g(x_*)$, such that

$$\langle \lambda_0 \nabla f(x_*) + \lambda \nabla g(x_*), x - x_* \rangle \ge 0, \quad \forall x \in X, \quad \lambda g(x_*) = 0.$$

Since λ_0 and λ are not equal to 0 simultaneously, three cases are possible.

1. $\lambda_0 = 0$ and $\lambda > 0$. Then, by optimality conditions, $g(x_*) = 0$ and $\langle \lambda \nabla g(x_*), \tilde{x} - x_* \rangle \ge 0$. Thus, by the Lemma assumption and strong convexity,

$$\varepsilon \ge g(\tilde{x}) \ge g(x_*) + \langle \nabla g(x_*), \tilde{x} - x_* \rangle + \frac{\mu}{2}\|\tilde{x} - x_*\|_E^2 \ge \frac{\mu}{2}\|\tilde{x} - x_*\|_E^2.$$

2. $\lambda_0 > 0$ and $\lambda = 0$. Then, by optimality conditions, $\langle \lambda_0 \nabla f(x_*), \tilde{x} - x_* \rangle \ge 0$. Thus, by the Lemma assumption and strong convexity,

$$f(x_*) + \varepsilon \ge f(\tilde{x}) \ge f(x_*) + \langle \nabla f(x_*), \tilde{x} - x_* \rangle + \frac{\mu}{2}\|\tilde{x} - x_*\|_E^2 \ge f(x_*) + \frac{\mu}{2}\|\tilde{x} - x_*\|_E^2.$$

3. $\lambda_0 > 0$, $\lambda > 0$. Then, by optimality conditions, $g(x_*) = 0$ and $\langle \lambda_0 \nabla f(x_*) + \lambda \nabla g(x_*), \tilde{x} - x_* \rangle \ge 0$. Thus, either $\langle \nabla g(x_*), \tilde{x} - x_* \rangle \ge 0$ and the proof is the same as in the item 1, or $\langle \nabla f(x_*), \tilde{x} - x_* \rangle \ge 0$ and the proof is the same as in the item 2. □

Theorem 3 *Assume that inequalities (8.5) and (8.7) hold and f, g are strongly convex with the same parameter μ. Also assume that the prox function $d(x)$ satisfies (8.18) and the starting point $x_0 \in X$ and a number $R_0 > 0$ are such that $\|x_0 - x_*\|_E^2 \le R_0^2$. Then, the point x_p returned by Algorithm 2 is an ε-solution to (8.1) in the sense of (8.6) and $\|x_p - x_*\|_E^2 \le \frac{2\varepsilon}{\mu}$. At the same time, the total number of iterations of Algorithm 1 does not exceed*

$$\left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil + \frac{32\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon}. \tag{8.19}$$

Proof Observe that, for all $p \ge 0$, the function $d_p(x)$ defined in Algorithm 2 is 1-strongly convex w.r.t. the norm $\|\cdot\|_E / R_p$. The conjugate of this norm is $R_p\|\cdot\|_{E,*}$.

Algorithm 2 Adaptive mirror descent (non-smooth strongly convex objective)

Input: accuracy $\varepsilon > 0$; strong convexity parameter μ; Ω s.t. $d(x) \leq \frac{\Omega}{2} \quad \forall x \in X : \|x\|_E \leq 1$; starting point x_0 and number R_0 s.t. $\|x_0 - x_*\|_E^2 \leq R_0^2$.

Set $d_0(x) = d\left(\frac{x-x_0}{R_0}\right)$.
Set $p = 1$.
repeat
    Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
    Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
    Set x_p as the output of Algorithm 1 with accuracy ε_p, prox-function $d_{p-1}(\cdot)$ and $\frac{\Omega}{2}$ as Θ_0^2.
    $d_p(x) \leftarrow d\left(\frac{x-x_p}{R_p}\right)$.
    Set $p = p + 1$.
until $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.
Output: x_p.

This means that, at each step k of inner Algorithm 1, M_k changes to $M_k R_{p-1}$, where $p \geq 1$ is the number of outer iteration.

We show, by induction, that, for all $p \geq 0$, $\|x_p - x_*\|_E^2 \leq R_p^2$. For $p = 0$ it holds by the assumption on x_0 and R_0. Let us assume that this inequality holds for some p and show that it holds for $p + 1$. By (8.18), we have $d_p(x_*) \leq \frac{\Omega}{2}$. Thus, on the outer iteration $p + 1$, by Theorem 1 and (8.6), after at most

$$k_{p+1} = \left\lceil \frac{\Omega \max\{M_f^2, M_g^2\} R_p^2}{\varepsilon_{p+1}^2} \right\rceil \tag{8.20}$$

inner iterations, $x_{p+1} = \bar{x}^{k_{p+1}}$ satisfies

$$f(x_{p+1}) - f(x_*) \leq \varepsilon_{p+1}, \quad g(x_{p+1}) \leq \varepsilon_{p+1},$$

where $\varepsilon_{p+1} = \frac{\mu R_{p+1}^2}{2}$. Then, by Lemma 2,

$$\|x_{p+1} - x_*\|_E^2 \leq \frac{2\varepsilon_{p+1}}{\mu} = R_{p+1}^2.$$

Thus, we proved that, for all $p \geq 0$, $\|x_p - x_*\|_E^2 \leq R_p^2 = R_0^2 \cdot 2^{-p}$. At the same time, we have, for all $p \geq 1$,

$$f(x_p) - f(x_*) \leq \frac{\mu R_0^2}{2} \cdot 2^{-p}, \quad g(x_p) \leq \frac{\mu R_0^2}{2} \cdot 2^{-p}.$$

Thus, if $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$, x_p is an ε-solution to (8.1) in the sense of (8.6) and

$$\|x_p - x_*\|_E^2 \leq R_0^2 \cdot 2^{-p} \leq \frac{2\varepsilon}{\mu}.$$

Let us now estimate the total number N of inner iterations, i.e., the iterations of Algorithm 1. Let us denote $\hat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$. According to (8.20), we have

$$\begin{aligned} N = \sum_{p=1}^{\hat{p}} k_p &\leq \sum_{p=1}^{\hat{p}} \left(1 + \frac{\Omega \max\{M_f^2, M_g^2\} R_p^2}{\varepsilon_{p+1}^2} \right) \\ &= \sum_{p=1}^{\hat{p}} \left(1 + \frac{16\Omega \max\{M_f^2, M_g^2\} 2^p}{\mu^2 R_0^2} \right) \\ &\leq \hat{p} + \frac{32\Omega \max\{M_f^2, M_g^2\} 2^{\hat{p}}}{\mu^2 R_0^2} \leq \hat{p} + \frac{32\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon}. \end{aligned}$$

□

Similarly to Sect. 8.3.1, let us consider a special type of problem (8.1) with strongly convex g given by

$$g(x) = \max_{i \in \{1,\ldots,m\}} \big\{ g_i(x) \big\}. \tag{8.21}$$

and corresponding dual problem

$$\varphi(\lambda) = \min_{x \in X} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right\} \to \max_{\lambda_i \geq 0, i \in \{1,\ldots,m\}} \varphi(\lambda).$$

On each outer iteration p of Algorithm 2, there is the last inner iteration k_p of Algorithm 1. We define approximate dual solution as $\lambda_p = \bar{\lambda}^{k_p}$, where $\bar{\lambda}^{k_p}$ is defined in (8.14). We modify Algorithm 2 to return a pair (x_p, λ_p).

Combining Theorems 2 and 3, we obtain the following result.

Theorem 4 *Assume that g is given by* (8.21)*, inequalities* (8.5) *and* (8.7) *hold and f, g are strongly convex with the same parameter μ. Also assume that the prox function $d(x)$ satisfies* (8.18) *and the starting point $x_0 \in X$ and a number $R_0 > 0$ are such that $\|x_0 - x_*\|_E^2 \leq R_0^2$. Then, the pair (x_p, λ_p) returned by Algorithm 2 satisfies*

$$f(x_p) - \varphi(\lambda_p) \leq \varepsilon, \quad g(x_p) \leq \varepsilon.$$

and $\|x_p - x_*\|_E^2 \le \frac{2\varepsilon}{\mu}$. *At the same time, the total number of inner iterations of Algorithm 1 does not exceed*

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

8.3.3 General Convex Objective Function

In this subsection, we assume that the objective function f in (8.1) might not satisfy (8.7) and, hence, its subgradients could be unbounded. One of the examples is a quadratic function. We also assume that inequality (8.8) holds.

We further develop ideas in [21, 24] and adapt them for problem (8.1), in a way that our algorithm allows to use non-Euclidean proximal setup, as does Mirror Descent, and does not require to know the constant M_g. Following [21], given a function f for each subgradient $\nabla f(x)$ at a point $y \in X$, we define

$$v_f[y](x) = \begin{cases} \left\langle \frac{\nabla f(x)}{\|\nabla f(x)\|_{E,*}}, x - y \right\rangle, & \nabla f(x) \neq 0 \\ 0 & \nabla f(x) = 0 \end{cases}, \quad x \in X. \tag{8.22}$$

The following result gives complexity estimate for Algorithm 3 in terms of $v_f[x_*](x)$. Below we use this theorem to establish complexity result for smooth objective f.

Theorem 5 *Assume that inequality* (8.5) *holds and a known constant* $\Theta_0 > 0$ *is such that* $d(x_*) \le \Theta_0^2$. *Then, Algorithm 3 stops after not more than*

$$k = \left\lceil \frac{2\max\{1, M_g^2\}\Theta_0^2}{\varepsilon^2} \right\rceil \tag{8.23}$$

iterations and it holds that $\min_{i \in I} v_f[x_*](x^i) \le \varepsilon$ *and* $g(\bar{x}^k) \le \varepsilon$.

Proof First, let us prove that the inequality in the stopping criterion holds for k defined in (8.23). Denote $[k] = \{i \in \{0, \dots, k-1\}\}$, $J = [k] \setminus I$. By (8.5), we have that, for any $j \in J$, $\|\nabla g(x^j)\|_{E,*} \le M_g$. Hence, since $|I| + |J| = k$, by (8.23), we obtain

$$|I| + \sum_{j \in J} \frac{1}{\|\nabla g(x^j)\|_{E,*}^2} \ge |I| + \frac{|J|}{M_g^2} \ge \frac{k}{\max\{1, M_g^2\}} \ge \frac{2\Theta_0^2}{\varepsilon^2}.$$

Algorithm 3 Adaptive mirror descent (general convex objective)

Input: accuracy $\varepsilon > 0$; Θ_0 s.t. $d(x_*) \leq \Theta_0^2$.

$x^0 = \arg\min_{x \in X} d(x)$.
Initialize the set I as empty set.
Set $k = 0$.
repeat
  if $g(x^k) \leq \varepsilon$ **then**
    $h_k = \frac{\varepsilon}{\|\nabla f(x^k)\|_{E,*}}$
    $x^{k+1} = \text{Mirr}[x^k](h_k \nabla f(x^k))$ ("productive step")
    Add k to I.
  else
    $h_k = \frac{\varepsilon}{\|\nabla g(x^k)\|_{E,*}^2}$
    $x^{k+1} = \text{Mirr}[x^k](h_k \nabla g(x^k))$ ("non-productive step")
  end if
  Set $k = k + 1$.
until $|I| + \sum_{j \in J} \frac{1}{\|\nabla g(x^j)\|_{E,*}^2} \geq \frac{2\Theta_0^2}{\varepsilon^2}$

Output: $\bar{x}^k := \arg\min_{x^j, j \in I} f(x^j)$

From Lemma 1 with $u = x_*$ and $\Delta = 0$, by the definition of h_i, $i \in I$, we have, for all $i \in I$,

$$\begin{aligned} \varepsilon v_f[x_*](x^i) &= \varepsilon \left\langle \frac{\nabla f(x^i)}{\|\nabla f(x^i)\|_{E,*}}, x^i - x_* \right\rangle = h_i \langle \nabla f(x^i), x^i - x_* \rangle \\ &\leq \frac{h_i^2}{2} \|\nabla f(x^i)\|_{E,*}^2 + V[x^i](x_*) - V[x^{i+1}](x_*) \\ &= \frac{\varepsilon^2}{2} + V[x^i](x_*) - V[x^{i+1}](x_*). \end{aligned} \tag{8.24}$$

Similarly, by the definition of h_i, $i \in J$, we have, for all $i \in J$,

$$\begin{aligned} \frac{\varepsilon(g(x^i) - g(x_*))}{\|\nabla g(x^i)\|_{E,*}^2} &= h_i \big(g(x^i) - g(x_*)\big) \leq \frac{h_i^2}{2} \|\nabla g(x^i)\|_{E,*}^2 + V[x^i](x_*) - V[x^{i+1}](x_*) \\ &= \frac{\varepsilon^2}{2\|\nabla g(x^i)\|_{E,*}^2} + V[x^i](x_*) - V[x^{i+1}](x_*). \end{aligned}$$

Whence, using that, for all $i \in J$, $g(x^i) - g(x_*) \geq g(x^i) > \varepsilon$, we have

$$-\frac{\varepsilon^2}{2\|\nabla g(x^i)\|_{E,*}^2} + V[x^i](x_*) - V[x^{i+1}](x_*) > 0. \tag{8.25}$$

Summing up inequalities (8.24) for $i \in I$ and applying (8.25) for $i \in J$, we obtain

$$\varepsilon|I| \min_{i \in I} v_f[x_*](x^i) \le \varepsilon \sum_{i \in I} v_f[x_*](x^i) < \frac{\varepsilon^2}{2} \cdot |I| + \Theta_0^2 - \sum_{i \in J} \frac{\varepsilon^2}{2\|\nabla g(x^i)\|_{E,*}^2},$$

where we also used that, by definition of x^0 and (8.8),

$$V[x^0](x_*) = d(x_*) - d(x^0) - \langle \nabla d(x^0), x_* - x^0 \rangle \le d(x_*) \le \Theta_0^2.$$

If the stopping criterion in Algorithm 3 is fulfilled, we get

$$\varepsilon|I| \min_{i \in I} v_f[x_*](x^i) < \varepsilon^2 |I|.$$

Since the inequality is strict, the set I is not empty and the output point $\bar{x}^k$ is correctly defined. Dividing both sides of the last inequality by $\varepsilon|I|$, we obtain the first statement of the Theorem. By definition of $\bar{x}^k$, it is obvious that $g(\bar{x}^k) \le \varepsilon$. □

To obtain the complexity of our algorithm in terms of the values of the objective function f, we define non-decreasing function

$$\omega(\tau) = \begin{cases} \max\limits_{x \in X}\{f(x) - f(x_*) : \|x - x_*\|_E \le \tau\} & \tau \ge 0, \\ 0 & \tau < 0. \end{cases} \tag{8.26}$$

and use the following lemma from [21].

Lemma 3 *Assume that f is a convex function. Then, for any $x \in X$,*

$$f(x) - f(x_*) \leqslant \omega(v_f[x_*](x)). \tag{8.27}$$

Corollary 1 *Assume that the objective function f in (8.1) is defined as $f(x) = \max_{i \in \{1,\dots,m\}} f_i(x)$, where f_i, $i = 1, \dots, m$ are differentiable with Lipschitz-continuous gradient*

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{E,*} \le L_i \|x - y\|_E \quad \forall x, y \in X, \quad i \in \{1, \dots, m\}. \tag{8.28}$$

Then $\bar{x}^k$ is $\widetilde{\varepsilon}$-solution to (8.1) in the sense of (8.6), where

$$\widetilde{\varepsilon} = \max\{\varepsilon, \varepsilon \max_{i=1,\dots,m} \|\nabla f_i(x_*)\|_{E,*} + \varepsilon^2 \max_{i=1,\dots,m} L_i/2\}.$$

Proof As it was shown in Theorem 5, $g(\bar{x}^k) \le \varepsilon$. It follows from (8.28) that

$$\begin{aligned} f_i(x) &\le f_i(x_*) + \langle \nabla f_i(x_*), x - x_* \rangle + \frac{1}{2} L_i ||x - x_*||_E^2 \\ &\le f_i(x_*) + \|\nabla f_i(x_*)\|_{E,*} \|x - x_*\|_E + \frac{1}{2} L_i ||x - x_*||_E^2, \quad i = 1, \dots, m. \end{aligned}$$

Whence, $\omega(\tau) \leq \tau \max_{i=1,\dots,m} \|\nabla f_i(x_*)\|_{E,*} + \frac{\tau^2 \max_{i=1,\dots,m} L_i}{2}$. By Lemma 3, non-decreasing property of ω and Theorem 5, we obtain

$$\begin{aligned} f(\bar{x}^k) - f(x_*) &= \min_{i\in I} f(x^i) - f(x_*) \leq \min_{i\in I} \omega(v_f[x_*](x^i)) \\ &\leq \omega(\min_{i\in I} v_f[x_*](x^i)) \leq \omega(\varepsilon) \\ &\leq \varepsilon \max_{i=1,\dots,m} \|\nabla f_i(x_*)\|_{E,*} + \frac{\varepsilon^2 \max_{i=1,\dots,m} L_i}{2}. \end{aligned}$$

□

8.4 Randomization for Constrained Problems

In this section, we consider randomized version of problem (8.1). This means that we still can use the value of the function $g(x)$ in an algorithm, but, instead of subgradients of f and g, we use their stochastic approximations. We combine the idea of switching subgradient method [25] and Stochastic Mirror Descent method introduced in [16]. More general case of stochastic optimization problems with expectation constraints is studied in [14]. We consider convex problems as long as strongly convex and, for each case, we have two types of algorithms. The first one allows to control expectation of the objective residual $f(\tilde{x}) - f(x_*)$ and inequality infeasibility $g(\tilde{x})$, where $\tilde{x}$ is the output of the algorithm. The second one allows to control probability of large deviation for these two quantities.

We introduce the following new assumptions. Given a point $x \in X$, we can calculate stochastic subgradients $\nabla f(x, \xi)$, $\nabla g(x, \zeta)$, where ξ, ζ are random vectors. These stochastic subgradients satisfy

$$\mathbb{E}\big[\nabla f(x, \xi)\big] = \nabla f(x) \in \partial f(x), \quad \mathbb{E}\big[\nabla g(x, \zeta)\big] = \nabla g(x) \in \partial g(x), \tag{8.29}$$

and

$$\|\nabla f(x, \xi)\|_{E,*} \leq M_f, \quad \|\nabla g(x, \zeta)\|_{E,*} \leq M_g, \quad \text{a.s. in } \xi, \zeta. \tag{8.30}$$

To motivate these assumptions, we consider the following example.

Example 1 ([3]) Consider Problem (8.1) with

$$f(x) = \frac{1}{2}\langle Ax, x\rangle,$$

where A is given $n \times n$ matrix, $X = S(1)$ being standard unit simplex, i.e., $X = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$, and

$$g(x) = \max_{i \in \{1,\dots,m\}} \left\{ \langle c_i, x \rangle \right\},$$

where $\left\{c_i\right\}_{i=1}^m$ are given vectors in $\mathbb{R}^n$.

Even if the matrix A is sparse, the gradient $\nabla f(x) = Ax$ is usually not. The exact computation of the gradient takes $O(n^2)$ arithmetic operations, which is expensive when n is large. In this setting, it is natural to use randomization to construct a stochastic approximation for $\nabla f(x)$. Let ξ be a random variable taking its values in $\{1, \dots, n\}$ with probabilities $(x_1, \dots, x_n)$ respectively. Let $A^{\langle i \rangle}$ denote the i-th column of the matrix A. Since $x \in S_n(1)$,

$$\begin{aligned} \mathbb{E}\big[A^{\langle \xi \rangle}\big] &= A^{\langle 1 \rangle} \underbrace{\mathbb{P}\big(\xi = 1\big)}_{x_1} + \dots + A^{\langle n \rangle} \underbrace{\mathbb{P}\big(\xi = n\big)}_{x_n} \\ &= A^{\langle 1 \rangle} x_1 + \dots + A^{\langle n \rangle} x_n = Ax. \end{aligned}$$

Thus, we can use $A^{\langle \xi \rangle}$ as stochastic subgradient, which can be calculated in $O(n)$ arithmetic operations.

8.4.1 Convex Objective Function, Control of Expectation

In this subsection, we consider convex optimization problem (8.1) in randomized setting described above. In this setting the output of the algorithm is random. Thus, we need to change the notion of approximate solution. Let x_* be a solution to (8.1). We say that a (random) point $\tilde{x} \in X$ is an *expected ε-solution* to (8.1) if

$$\mathbb{E} f(\tilde{x}) - f(x_*) \le \varepsilon, \quad \text{and} \;\; g(\tilde{x}) \; \le \varepsilon \quad \text{a.s.} \tag{8.31}$$

We also introduce a stronger assumption than (8.8). Namely, we assume that we know a constant $\Theta_0 > 0$ such that

$$\sup_{x,y \in X} V[x](y) \le \Theta_0^2. \tag{8.32}$$

The main difference between the method, which we describe below, and the method in [14] is the adaptivity of our method both in terms of stepsize and stopping rule, which means that we do not need to know the constants M_f, M_g in advance. We assume that on each iteration of the algorithm independent realizations of ξ and ζ are generated. The algorithm is similar to the one in [3], but, for the sake of consistency with other parts of the chapter, we use slightly different proof.

Algorithm 4 Adaptive stochastic mirror descent

Input: accuracy $\varepsilon > 0$; Θ_0 s.t. $V[x](y) \le \Theta_0^2, \quad \forall x, y \in X$.

$x^0 = \arg\min_{x \in X} d(x)$.
Initialize the set I as empty set.
Set $k = 0$.
repeat
  if $g(x^k) \le \varepsilon$. **then**
    $M_k = \|\nabla f(x^k, \xi^k)\|_{E,*}$.
    $h_k = \Theta_0 \Big(\sum_{i=0}^{k} M_i^2 \Big)^{-1/2}$.
    $x^{k+1} = \mathrm{Mirr}[x^k](h_k \nabla f(x^k, \xi^k))$ ("productive step").
    Add k to I.
  else
    $M_k = \|\nabla g(x^k, \zeta^k)\|_{E,*}$.
    $h_k = \Theta_0 \Big(\sum_{i=0}^{k} M_i^2 \Big)^{-1/2}$.
    $x^{k+1} = \mathrm{Mirr}[x^k](h_k \nabla g(x^k, \zeta^k))$ ("non-productive step").
  end if
  Set $k = k + 1$.
until $k \ge \frac{2\Theta_0}{\varepsilon} \Big(\sum_{i=0}^{k-1} M_i^2 \Big)^{1/2}$.

Output: $\bar{x}^k = \frac{1}{|I|} \sum_{k \in I} x^k$.

Theorem 6 *Let equalities (8.29) and inequalities (8.30) hold. Assume that a known constant* $\Theta_0 > 0$ *is such that* $V[x](y) \le \Theta_0^2, \quad \forall x, y \in X$. *Then, Algorithm 4 stops after not more than*

$$k = \left\lceil \frac{4 \max\{M_f^2, M_g^2\} \Theta_0^2}{\varepsilon^2} \right\rceil \tag{8.33}$$

iterations and $\bar{x}^k$ *is an expected* ε*-solution to (8.1) in the sense of (8.31).*

Proof First, let us prove that the inequality in the stopping criterion holds for k defined in (8.33). By (8.30), we have that, for any $i \in \{0, \ldots, k-1\}$, $M_i \le \max\{M_f, M_g\}$. Hence, by (8.33), $\frac{2\Theta_0}{\varepsilon} \left(\sum_{j=0}^{k-1} M_j^2 \right)^{1/2} \le \frac{2\Theta_0}{\varepsilon} \max\{M_f, M_g\} \sqrt{k} \le k$.

Denote $[k] = \{i \in \{0, \ldots, k-1\}\}$, $J = [k] \setminus I$ and

$$\delta_i = \begin{cases} \langle \nabla f(x^i, \xi^i) - \nabla f(x^i), x_* - x^i \rangle, & \text{if } i \in I, \\ \langle \nabla g(x^i, \zeta^i) - \nabla g(x^i), x_* - x^i \rangle, & \text{if } i \in J. \end{cases} \tag{8.34}$$

From Lemma 1 with $u = x_*$ and $\Delta = \nabla f(x^i, \xi^i) - \nabla f(x^i)$, we have, for all $i \in I$,

$$h_i\big(f(x^i) - f(x_*)\big) \le \frac{h_i^2}{2}\|\nabla f(x^i, \xi^i)\|_{E,*}^2 + V[x^i](x_*) - V[x^{i+1}](x_*) + h_i\delta_i$$

and, from Lemma 1 with $u = x_*$ and $\Delta = \nabla g(x^i, \zeta^i) - \nabla g(x^i)$, for all $i \in J$,

$$h_i\big(g(x^i) - g(x_*)\big) \le \frac{h_i^2}{2}\|\nabla g(x^i, \zeta^i)\|_{E,*}^2 + V[x^i](x_*) - V[x^{i+1}](x_*) + h_i\delta_i.$$

Dividing each inequality by h_i and summing up these inequalities for i from 0 to $k-1$, using the definition of h_i, $i \in \{0, \dots, k-1\}$, we obtain

$$\sum_{i\in I}\big(f(x^i) - f(x_*)\big) + \sum_{i\in J}\big(g(x^i) - g(x_*)\big)$$
$$\le \sum_{i\in[k]}\frac{h_i M_i^2}{2} + \sum_{i\in[k]}\frac{1}{h_i}\big(V[x^i](x_*) - V[x^{i+1}](x_*)\big) + \sum_{i\in[k]}\delta_i \tag{8.35}$$

Using (8.32), we get

$$\sum_{i=0}^{k-1}\frac{1}{h_i}\big(V[x^i](x_*) - V[x^{i+1}](x_*)\big)$$
$$= \frac{1}{h_0}V[x^0](x_*) + \sum_{i=0}^{k-2}\left(\frac{1}{h_{i+1}} - \frac{1}{h_i}\right)V[x^{i+1}](x_*) - \frac{1}{h_{k-1}}V[x^k](x_*)$$
$$\le \frac{\Theta_0^2}{h_0} + \Theta_0^2\sum_{k=0}^{k-2}\left(\frac{1}{h_{i+1}} - \frac{1}{h_i}\right) = \frac{\Theta_0^2}{h_{k-1}}.$$

Whence, by the definition of stepsizes h_i,

$$\sum_{i\in I}\big(f(x^i) - f(x_*)\big) + \sum_{i\in J}\big(g(x^i) - g(x_*)\big) \le \sum_{i\in[k]}\frac{h_i M_i^2}{2} + \frac{\Theta_0^2}{h_{k-1}} + \sum_{i\in[k]}\delta_i$$
$$\le \sum_{i=0}^{k-1}\frac{\Theta_0}{2}\frac{M_i^2}{\left(\sum_{j=0}^{i} M_j^2\right)^{1/2}} + \Theta_0\left(\sum_{i=0}^{k-1}M_i^2\right)^{1/2} + \sum_{i\in[k]}\delta_i$$
$$\le 2\Theta_0\left(\sum_{i=0}^{k-1}M_i^2\right)^{1/2} + \sum_{i\in[k]}\delta_i,$$

where we used inequality $\sum_{i=0}^{k-1} \frac{M_i^2}{\left(\sum_{j=0}^{i} M_j^2\right)^{1/2}} \le 2\left(\sum_{i=0}^{k-1} M_i^2\right)^{1/2}$, which can be proved by induction. Since, for $i \in J$, $g(x^i) - g(x_*) \ge g(x^i) > \varepsilon$, by convexity of f, the definition of $\bar{x}^k$, and the stopping criterion, we get

$$|I|\big(f(\bar{x}^k) - f(x_*)\big) < \varepsilon|I| - \varepsilon k + 2\Theta_0\Big(\sum_{i=0}^{k-1} M_i^2\Big)^{1/2} + \sum_{i=0}^{k-1}\delta_i \le \varepsilon|I| + \sum_{i=0}^{k-1}\delta_i. \tag{8.36}$$

Taking the expectation and using (8.29), as long as the inequality is strict and the case of $I = \emptyset$ is impossible, we obtain

$$\mathbb{E} f(\bar{x}^k) - f(x_*) \le \varepsilon. \tag{8.37}$$

At the same time, for $i \in I$ it holds that $g(x^i) \le \varepsilon$. Then, by the definition of $\bar{x}^k$ and the convexity of g,

$$g(\bar{x}^k) \le \frac{1}{|I|}\sum_{i\in I} g(x^i) \le \varepsilon.$$

□

8.4.2 Convex Objective Function, Control of Large Deviation

In this subsection, we consider the same setting as in previous subsection, but change the notion of approximate solution. Let x_* be a solution to (8.1). Given $\varepsilon > 0$ and $\sigma \in (0, 1)$, we say that a point $\tilde{x} \in X$ is an (ε, σ)-*solution* to (8.1) if

$$\mathbb{P}\{f(\tilde{x}) - f(x_*) \le \varepsilon, \quad g(\tilde{x}) \le \varepsilon\} \ge 1 - \sigma. \tag{8.38}$$

As in the previous subsection, we use an assumption expressed by inequality (8.32). We assume additionally to (8.30) that inequalities (8.5) and (8.7) hold. Unfortunately, it is not clear, how to obtain large deviation guarantee for an adaptive method. Thus, in this section, we assume that the constants M_f, M_g are known and use a simplified algorithm. We assume that on each iteration of the algorithm independent realizations of ξ and ζ are generated.

To analyze Algorithm 5 in terms of large deviation bound, we need the following known result, see, e.g., [10].

Lemma 4 (Azuma-Hoeffding Inequality) *Let $\eta^1, \dots, \eta^n$ be a sequence of independent random variables taking their values in some set Ξ, and let $Z =$*

Algorithm 5 Stochastic mirror descent

Input: accuracy $\varepsilon > 0$; maximum number of iterations N; M_f, M_g s.t. (8.5), (8.7), (8.30) hold.

$x^0 = \arg\min_{x \in X} d(x)$.
Set $h = \frac{\varepsilon}{\max\{M_f^2, M_g^2\}}$.
Set $k = 0$.
repeat
  if $g(x^k) \le \varepsilon$. **then**
    $x^{k+1} = \mathrm{Mirr}[x^k](h\nabla f(x^k, \xi^k))$ ("productive step").
    Add k to I.
  else
    $x^{k+1} = \mathrm{Mirr}[x^k](h\nabla g(x^k, \zeta^k))$ ("non-productive step").
  end if
  Set $k = k + 1$.
until $k \ge N$.

Output: If $I \neq \emptyset$, then $\bar{x}^k = \frac{1}{|I|}\sum_{k \in I} x^k$. Otherwise $\bar{x}^k = NULL$.

$\phi(\eta^1, \ldots, \eta^n)$ *for some function* $\phi : \Xi^n \to \mathbb{R}$. *Suppose that a. s.*

$$\left|\mathbb{E}[Z|\eta^1, \ldots, \eta^i] - \mathbb{E}[Z|\eta^1, \ldots, \eta^{i-1}]\right| \le c_i, \quad i = 1, \ldots, n,$$

where c_i, $i \in \{1, \ldots, n\}$ *are deterministic. Then, for each* $t \ge 0$

$$\mathbb{P}\big(Z - \mathbb{E}Z \ge t\big) \le \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right\}.$$

Theorem 7 *Let equalities* (8.29) *and inequalities* (8.5), (8.7), (8.30) *hold. Assume that a known constant* $\Theta_0 > 0$ *is such that* $V[x](y) \le \Theta_0^2$, $\forall x, y \in X$, *and the confidence level satisfies* $\sigma \in (0, 0.5)$. *Then, if in Algorithm* 5

$$N = \left\lceil 70 \frac{\max\{M_f^2, M_g^2\}\Theta_0^2}{\varepsilon^2} \ln\frac{1}{\sigma} \right\rceil, \tag{8.39}$$

$\bar{x}^k$ *is an* (ε, σ)*-solution to* (8.1) *in the sense of* (8.38).

Proof Let us denote $M = \max\{M_f, M_g\}$. In the same way as we obtained (8.35) in the proof of Theorem 6, we obtain

$$h\sum_{i \in I}\big(f(x^i) - f(x_*)\big) + h\sum_{i \in J}\big(g(x^i) - g(x_*)\big)$$

$$\le \frac{h^2 M^2 k}{2} + V[x^0](x_*) + h\sum_{i=0}^{k-1}\delta_i,$$

where δ_i, $i = 0, \dots, k-1$ are defined in (8.34). Since, for $i \in J$, $g(x^i) - g(x_*) \geq g(x^i) > \varepsilon$, by convexity of f, the definition of $\bar{x}^k$ and h, we get

$$h|I|\big(f(\bar{x}^k) - f(x_*)\big) < \varepsilon h|I| - \frac{\varepsilon^2 k}{2M^2} + \Theta_0^2 + h\sum_{i=0}^{k-1}\delta_i. \tag{8.40}$$

Using Cauchy-Schwarz inequality, (8.5), (8.7), (8.30), (8.32), we have

$$\begin{aligned} h\big|\delta_i\big| &\leq 2hM\|x^i - x^*\| \\ &\leq 2hM\sqrt{2V[x^i](x^*)} \leq 2\sqrt{2}hM\Theta_0 = 2\sqrt{2}\frac{\varepsilon\Theta_0}{M}. \end{aligned}$$

Now we use Lemma 4 with $Z = \sum\limits_{i=0}^{k-1} h\delta_i$. Clearly, $\mathbb{E}Z = \mathbb{E}\Big[\sum\limits_{i=0}^{k-1} h\delta_i\Big] = 0$ and we can take $c_i = 2\sqrt{2}\frac{\varepsilon\Theta_0}{M}$. Then, by Lemma 4, for each $t \geq 0$,

$$\mathbb{P}\left\{\sum_{i=0}^{k-1} h\delta_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\sum\limits_{i=0}^{k-1} c_i^2}\right) = \exp\left(-\frac{t^2M^2}{16\varepsilon^2\Theta_0^2 k}\right).$$

In other words, for each $\sigma \in (0, 1)$

$$\mathbb{P}\left\{\sum_{i=0}^{k-1} h\delta_i \geq \frac{4\varepsilon\Theta_0}{M}\sqrt{k\ln\left(\frac{1}{\sigma}\right)}\right\} \leq \sigma.$$

Applying this inequality to (8.40), we obtain, for any $\sigma \in (0, 1)$,

$$\mathbb{P}\left\{h|I|\big(f(\bar{x}^k) - f(x_*)\big) < \varepsilon h|I| - \frac{\varepsilon^2 k}{2M^2} + \Theta_0^2 + \frac{4\varepsilon\Theta_0}{M}\sqrt{k\ln\left(\frac{1}{\sigma}\right)}\right\} \geq 1 - \sigma.$$

Then, by (8.39), we have

$$\begin{aligned} -\frac{\varepsilon^2 k}{2M^2} + \Theta_0^2 + \frac{4\varepsilon\Theta_0}{M}\sqrt{k\ln\left(\frac{1}{\sigma}\right)} &< \Theta_0^2\left(-\frac{71}{2}\ln\left(\frac{1}{\sigma}\right) + 1 + 4\ln\left(\frac{1}{\sigma}\right)\sqrt{71}\right) \\ &< \Theta_0^2\left(-\frac{3}{2}\ln\left(\frac{1}{\sigma}\right) + 1\right). \end{aligned} \tag{8.41}$$

Since $\sigma \le 0.5 < \exp(-2/3)$, we have $-\frac{3}{2}\ln\left(\frac{1}{\sigma}\right) + 1 < 0$ and

$$\mathbb{P}\left\{h|I|\big(f(\bar{x}^k) - f(x^*)\big) < h|I|\varepsilon\right\} \ge 1 - \sigma.$$

Thus, with probability at least $1-\sigma$, the inequality is strict, the case of $I = \emptyset$ is impossible, and $\bar{x}^k$ is correctly defined. Dividing the both sides of it by $h \cdot |I|$, we obtain that $\mathbb{P}\left\{f(\bar{x}^k) - f(x^*) \le \varepsilon\right\} \ge 1 - \sigma$. At the same time, for $i \in I$ it holds that $g(x^i) \le \varepsilon$. Then, by the definition of $\bar{x}^k$ and the convexity of g, again with probability at least $1-\sigma$

$$g(\bar{x}^k) \le \frac{1}{|I|}\sum_{i \in I} g(x^i) \le \varepsilon.$$

Thus, $\bar{x}^k$ is an (ε, σ)-solution to (8.1) in the sense of (8.38). □

8.4.3 Strongly Convex Objective Function, Control of Expectation

In this subsection, we consider the setting of Sect. 8.4.1, but, as in Sect. 8.3.2, make the following additional assumptions. First, we assume that functions f and g are strongly convex. Second, without loss of generality, we assume that $0 = \arg\min_{x \in X} d(x)$. Third, we assume that we are given a starting point $x_0 \in X$ and a number $R_0 > 0$ such that $\|x_0 - x_*\|_E^2 \le R_0^2$. Finally, we make the following assumption (cf. (8.18)) that d is bounded in the following sense. Assume that x_* is some fixed point and x is a random point such that $\mathbb{E}_x\left[\|x - x_*\|_E^2\right] \le R^2$, then

$$\mathbb{E}_x\left[d\left(\frac{x - x_*}{R}\right)\right] \le \frac{\Omega}{2}, \tag{8.42}$$

where Ω is some known number and $\mathbb{E}_x$ denotes the expectation with respect to random vector x. For example, this assumption holds for Euclidean proximal setup. Unlike the method introduced in [14] for strongly convex problems, we present a method, which is based on the restart of Algorithm 5. Unfortunately, it is not clear, whether the restart technique can be combined with adaptivity to constants M_f, M_g. Thus, we assume that these constants are known.

The following lemma can be proved in the same way as Lemma 2.

Lemma 5 *Let f and g be strongly convex functions with the same parameter μ and x_* be a solution of problem (8.1). Assume that, for some random $\tilde{x} \in X$,*

$$\mathbb{E}f(\tilde{x}) - f(x_*) \le \varepsilon, \quad g(\tilde{x}) \le \varepsilon.$$

Then

$$\frac{\mu}{2}\mathbb{E}\|\tilde{x} - x_*\|_E^2 \leq \varepsilon.$$

Theorem 8 *Let equalities* (8.29) *and inequalities* (8.30) *hold and* f, g *be strongly convex with the same parameter* μ. *Also assume that the prox function* $d(x)$ *satisfies* (8.42) *and the starting point* $x_0 \in X$ *and a number* $R_0 > 0$ *are such that* $\|x_0 - x_*\|_E^2 \leq R_0^2$. *Then, the point* x_p *returned by Algorithm 6 is an expected* ε*-solution to* (8.1) *in the sense of* (8.31) *and* $\mathbb{E}\|x_p - x_*\|_E^2 \leq \frac{2\varepsilon}{\mu}$. *At the same time, the total number of inner iterations of Algorithm 5 does not exceed*

$$\left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil + \frac{32\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon}. \tag{8.43}$$

Proof Let us denote $M = \max\{M_f, M_g\}$. Observe that, for all $p \geq 0$, the function $d_p(x)$ defined in Algorithm 6 is 1-strongly convex w.r.t. the norm $\|\cdot\|_E/R_p$. The conjugate of this norm is $R_p\|\cdot\|_{E,*}$. This means that, at each outer iteration p, M changes to MR_{p-1}, where p is the number of outer iteration. We show by induction that, for all $p \geq 0$, $\mathbb{E}\|x_p - x_*\|_E^2 \leq R_p^2$. For $p = 0$ it holds by the definition of x_0 and R_0.

Let us assume that this inequality holds for some $p - 1$ and show that it holds for p. At iteration p, we start Algorithm 5 with starting point x_{p-1} and stepsize $h_p = \frac{\varepsilon_p}{M^2 R_{p-1}^2}$. Using the same steps as in the proof of Theorem 7, after N_p iterations

---

Algorithm 6 Stochastic mirror descent (strongly convex objective, expectation control)

Input: accuracy $\varepsilon > 0$; strong convexity parameter μ; Ω s.t. $\mathbb{E}_x\left[d\left(\frac{x-x_*}{R}\right)\right] \leq \frac{\Omega}{2}$ if $\mathbb{E}_x\left[\|x - x_*\|_E^2\right] \leq R^2$; starting point x_0 and number R_0 s.t. $\|x_0 - x_*\|_E^2 \leq R_0^2$.

Set $d_0(x) = d\left(\frac{x-x_0}{R_0}\right)$.

Set $p = 1$.

repeat

  Set $R_p^2 = R_0^2 \cdot 2^{-p}$.

  Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.

  Set $N_p = \left\lceil \frac{\max\{M_f^2, M_g^2\}\Omega R_{p-1}^2}{\varepsilon_p^2} \right\rceil$

  Set x_p as the output of Algorithm 5 with accuracy ε_p, number of iterations N_p, prox-function $d_{p-1}(\cdot)$ and $\frac{\Omega}{2}$ as Θ_0^2.

  $d_p(x) \leftarrow d\left(\frac{x-x_p}{R_p}\right)$.

  Set $p = p + 1$.

until $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Output: x_p.

---

of Algorithm 5 (see (8.40)), we obtain

$$h_p|I_p|\big(f(\bar{x}_p^k) - f(x_*)\big) < \varepsilon_p h_p |I_p| - \frac{\varepsilon_p^2 N_p}{2M^2 R_{p-1}^2} + V_{p-1}[x_{p-1}](x_*) + h_p \sum_{i=0}^{N_p-1} \delta_i, \tag{8.44}$$

where $V_{p-1}[z](x)$ is the Bregman divergence corresponding to $d_{p-1}(x)$ and I_p is the set of "productive steps". Using the definition of d_{p-1}, we have

$$V_{p-1}[x_{p-1}](x_*) = d_{p-1}(x_*) - d_{p-1}(x_{p-1}) - \langle \nabla d_{p-1}(x_{p-1}), x_* - x_{p-1} \rangle \le d_{p-1}(x_*).$$

Taking expectation with respect to x_{p-1} in (8.44) and using inductive assumption $\mathbb{E}\|x_{p-1} - x_*\|_E^2 \le R_{p-1}^2$ and (8.42), we obtain, substituting N_p,

$$\begin{aligned} h_p|I_p|\big(f(\bar{x}_p^k) - f(x_*)\big) &< \varepsilon_p h_p |I_p| - \frac{\varepsilon_p^2 N_p}{2M^2 R_{p-1}^2} + \frac{\Omega}{2} + h_p \sum_{i=0}^{N_p-1} \delta_i \\ &\le \varepsilon_p h_p |I_p| + h_p \sum_{i=0}^{N_p-1} \delta_i. \end{aligned} \tag{8.45}$$

Taking the expectation and using (8.29), as long as the inequality is strict and the case of $I_p = \emptyset$ is impossible, we obtain

$$\mathbb{E} f(\bar{x}_p^k) - f(x_*) \le \varepsilon_p. \tag{8.46}$$

At the same time, for $i \in I_p$ it holds that $g(x^i) \le \varepsilon_p$. Then, by the definition of $\bar{x}_p^k$ and the convexity of g,

$$g(\bar{x}_p^k) \le \frac{1}{|I_p|} \sum_{i \in I_p} g(x^i) \le \varepsilon_p.$$

Thus, we can apply Lemma 5 and obtain

$$\mathbb{E}\|x_p - x_*\|_E^2 \le \frac{2\varepsilon_p}{\mu} = R_p^2.$$

Thus, we proved that, for all $p \ge 0$, $\mathbb{E}\|x_p - x_*\|_E^2 \le R_p^2 = R_0^2 \cdot 2^{-p}$. At the same time, we have, for all $p \ge 1$,

$$\mathbb{E} f(x_p) - f(x_*) \le \frac{\mu R_0^2}{2} \cdot 2^{-p}, \quad g(x_p) \le \frac{\mu R_0^2}{2} \cdot 2^{-p}.$$

Thus, if $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$, x_p is an ε-solution to (8.1) in the sense of (8.31) and

$$\mathbb{E}\|x_p - x_*\|_E^2 \le R_0^2 \cdot 2^{-p} \le \frac{2\varepsilon}{\mu}.$$

Let us now estimate the total number N of inner iterations, i.e., the iterations of Algorithm 1. Let us denote $\hat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$. We have

$$N = \sum_{p=1}^{\hat{p}} N_p \le \sum_{p=1}^{\hat{p}} \left(1 + \frac{\Omega \max\{M_f^2, M_g^2\} R_{p-1}^2}{\varepsilon_p^2} \right)$$
$$= \sum_{p=1}^{\hat{p}} \left(1 + \frac{16\Omega \max\{M_f^2, M_g^2\} 2^p}{\mu^2 R_0^2} \right) \le \hat{p} + \frac{32\Omega \max\{M_f^2, M_g^2\} 2^{\hat{p}}}{\mu^2 R_0^2}$$
$$\le \hat{p} + \frac{32\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon}.$$

□

8.4.4 Strongly Convex Objective Function, Control of Large Deviation

In this subsection, we consider the setting of Sect. 8.4.2, but make the following additional assumptions. First, we assume that functions f and g are strongly convex. Second, without loss of generality, we assume that $0 = \arg\min_{x \in X} d(x)$. Third, we assume that we are given a starting point $x_0 \in X$ and a number $R_0 > 0$ such that $\|x_0 - x_*\|_E^2 \le R_0^2$. Finally, instead of (8.32), we assume that the Bregman divergence satisfies quadratic growth condition

$$V[z](x) \le \frac{\Omega}{2} \|x - z\|_E^2, \quad x, z \in X. \tag{8.47}$$

where Ω is some known number. For example, this assumption holds for Euclidean proximal setup. Unlike the method introduced in [14] for strongly convex problems, we present a method, which is based on the restart of Algorithm 5. Unfortunately, it is not clear, whether the restart technique can be combined with adaptivity to constants M_f, M_g. Thus, we assume that these constants are known.

Theorem 9 *Let equalities* (8.29) *and inequalities* (8.5), (8.7), (8.30) *hold. Let* f, g *be strongly convex with the same parameter* μ. *Also assume that the Bregman divergence* $V[z](x)$ *satisfies* (8.47) *and the starting point* $x_0 \in X$ *and a number*

$R_0 > 0$ *are such that* $\|x_0 - x_*\|_E^2 \le R_0^2$. *Then, the point* x_p *returned by Algorithm 7 is an* (ε, σ)*-solution to* (8.1) *in the sense of* (8.38) *and* $\|x_p - x_*\|_E^2 \le \frac{2\varepsilon}{\mu}$ *with probability at least* $1 - \sigma$. *At the same time, the total number of inner iterations of Algorithm 5 does not exceed*

$$\left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil + \frac{2240\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon} \left(\ln \frac{1}{\sigma} + \ln \log_2 \frac{\mu R_0^2}{2\varepsilon} \right).$$

Proof Let us denote $M = \max\{M_f, M_g\}$. Observe that, for all $p \ge 0$, the function $d_p(x)$ defined in Algorithm 7 is 1-strongly convex w.r.t. the norm $\|\cdot\|_E / R_p$. The conjugate of this norm is $R_p \|\cdot\|_{E,*}$. This means that, at each outer iteration p, M changes to $M R_{p-1}$, where p is the number of outer iteration.

Let A_p, $p \ge 0$ be the event $A_p = \{\|x_p - x_*\|_E^2 \le R_p^2\}$ and $\bar{A}_p$ be its complement. Note that, by the definition of x_0 and R_0, A_0 holds with probability 1. Denote $\hat{p} = \left\lceil \log_2 \frac{\mu R_0^2}{2\varepsilon} \right\rceil$.

We now show by induction that, for all $p \ge 1$, $\mathbb{P}\{A_p | A_{p-1}\} \ge 1 - \frac{\sigma}{\hat{p}}$. By inductive assumption, A_{p-1} holds and we have $\|x_{p-1} - x_*\|_E^2 \le R_{p-1}^2$. At iteration p, we start Algorithm 5 with starting point x_{p-1}, feasible set X_p and Bregman

---

Algorithm 7 Stochastic mirror descent (strongly convex objective, control of large deviation)

Input: accuracy $\varepsilon > 0$; strong convexity parameter μ; Ω s.t. $V[x](y) \le \frac{\Omega}{2}\|x - y\|_E^2, \quad x, y \in X$; starting point x_0 and number R_0 s.t. $\|x_0 - x_*\|_E^2 \le R_0^2$.

Set $d_0(x) = d\left(\frac{x - x_0}{R_0}\right)$.
Set $p = 1$.
repeat
  Set $R_p^2 = R_0^2 \cdot 2^{-p}$.
  Set $\varepsilon_p = \frac{\mu R_p^2}{2}$.
  Set $N_p = \left\lceil 70 \frac{\max\{M_f^2, M_g^2\} \Omega R_{p-1}^2}{\varepsilon_p^2} \ln\left(\frac{1}{\sigma} \log_2 \frac{\mu R_0^2}{2\varepsilon}\right) \right\rceil$.
  Set $X_p = \{x \in X : \|x - x_{p-1}\|_E^2 \le R_{p-1}^2\}$.
  Set x_p as the output of Algorithm 5 with accuracy ε_p, number of iteration N_p, prox-function $d_{p-1}(\cdot)$, Ω as Θ_0^2 and X_p as the feasible set.
  $d_p(x) \leftarrow d\left(\frac{x - x_p}{R_p}\right)$.
  Set $p = p + 1$.
until $p > \log_2 \frac{\mu R_0^2}{2\varepsilon}$.

Output: x_p.

---

divergence $V_{p-1}[z](x)$ corresponding to $d_{p-1}(x)$. Thus, by (8.47), we have

$$\max_{x,z\in X_p} V_{p-1}[z](x) = \max_{x,z\in X_p} d\left(\frac{x-x_{p-1}}{R_{p-1}}\right) - d\left(\frac{z-x_{p-1}}{R_{p-1}}\right)$$
$$-\left\langle \nabla d\left(\frac{z-x_{p-1}}{R_{p-1}}\right), \frac{x-x_{p-1}}{R_{p-1}} - \frac{z-x_{p-1}}{R_{p-1}}\right\rangle$$
$$= \max_{x,z\in X_p} V\left[\frac{z-x_{p-1}}{R_{p-1}}\right]\left(\frac{x-x_{p-1}}{R_{p-1}}\right)$$
$$\leq \max_{x,z\in X_p} \frac{\Omega\|x-z\|_E^2}{2R_{p-1}^2} \leq \Omega.$$

Hence, by Theorem 7 with $\sigma_p = \frac{\sigma}{\hat{p}}$, after N_p iterations of Algorithm 5, we have

$$\mathbb{P}\left\{f(x_p) - f(x_*) \leq \varepsilon_p, \quad g(x_p) \leq \varepsilon_p | A_{p-1}\right\} \geq 1 - \frac{\sigma}{\hat{p}}.$$

Whence, by Lemma 2,

$$\mathbb{P}\left\{A_p | A_{p-1}\right\} = \mathbb{P}\left\{\|x_p - x_*\|_E^2 \leq R_p^2 | A_{p-1}\right\} \geq 1 - \frac{\sigma}{\hat{p}},$$

which finishes the induction proof.

At the same time,

$$\begin{aligned}
&\mathbb{P}\left\{f(x_{\hat{p}}) - f(x_*) > \varepsilon_{\hat{p}} \quad \text{or} \quad g(x_{\hat{p}}) > \varepsilon_{\hat{p}}\right\} \\
&= \mathbb{P}\left\{f(x_{\hat{p}}) - f(x_*) > \varepsilon_{\hat{p}} \quad \text{or} \quad g(x_{\hat{p}}) > \varepsilon_{\hat{p}} \middle| A_{\hat{p}-1} \cup \bar{A}_{\hat{p}-1}\right\} \\
&= \mathbb{P}\left\{f(x_{\hat{p}}) - f(x_*) > \varepsilon_{\hat{p}} \quad \text{or} \quad g(x_{\hat{p}}) > \varepsilon_{\hat{p}} \middle| A_{\hat{p}-1}\right\} \mathbb{P}\{A_{\hat{p}-1}\} \\
&\quad + \mathbb{P}\left\{f(x_{\hat{p}}) - f(x_*) > \varepsilon_{\hat{p}} \quad \text{or} \quad g(x_{\hat{p}}) > \varepsilon_{\hat{p}} \middle| \bar{A}_{\hat{p}-1}\right\} \mathbb{P}\{\bar{A}_{\hat{p}-1}\} \\
&\leq \frac{\sigma}{\hat{p}} + \mathbb{P}\{\bar{A}_{\hat{p}-1}\} \overset{(*)}{\leq} \frac{\sigma}{\hat{p}} + \mathbb{P}\left\{f(x_{\hat{p}-1}) - f(x_*) > \varepsilon_{\hat{p}-1} \quad \text{or} \quad g(x_{\hat{p}-1}) > \varepsilon_{\hat{p}-1}\right\} \\
&\leq 2\cdot\frac{\sigma}{\hat{p}} + \mathbb{P}\{\bar{A}_{\hat{p}-2}\} \leq \ldots \leq \frac{\hat{p}-1}{\hat{p}}\cdot\sigma + \mathbb{P}\{\bar{A}_1\},
\end{aligned} \tag{8.48}$$

where $(*)$ follows from Lemma 2. Using that $\mathbb{P}\{A_1\} = \mathbb{P}\{A_1|A_0\} \geq 1 - \frac{\sigma}{\hat{p}}$ and, hence, $\mathbb{P}\{\bar{A}_1\} \leq \frac{\sigma}{\hat{p}}$, we obtain

$$\mathbb{P}\left\{f(x_{\hat{p}}) - f(x_*) \leq \varepsilon, \quad g(x_{\hat{p}}) \leq \varepsilon\right\} \geq 1 - \sigma.$$

Hence,

$$\mathbb{P}\left\{\|x_{\hat{p}} - x_*\|_E^2 \le \frac{2\varepsilon}{\mu}\right\} \ge 1 - \sigma.$$

Let us now estimate the total number N of inner iterations, i.e., the iterations of Algorithm 5. We have

$$\begin{aligned}
N = \sum_{p=1}^{\hat{p}} N_p &\le \sum_{p=1}^{\hat{p}} \left(1 + 70\frac{\Omega \max\{M_f^2, M_g^2\} R_{p-1}^2}{\varepsilon_p^2} \ln\left(\frac{1}{\sigma}\log_2 \frac{\mu R_0^2}{2\varepsilon}\right)\right) \\
&= \sum_{p=1}^{\hat{p}} \left(1 + 1120\frac{\Omega \max\{M_f^2, M_g^2\} 2^p}{\mu^2 R_0^2} \ln\left(\frac{1}{\sigma}\log_2 \frac{\mu R_0^2}{2\varepsilon}\right)\right) \\
&\le \hat{p} + 2240\frac{\Omega \max\{M_f^2, M_g^2\} 2^{\hat{p}}}{\mu^2 R_0^2} \ln\left(\frac{1}{\sigma}\log_2 \frac{\mu R_0^2}{2\varepsilon}\right) \\
&\le \hat{p} + 2240\frac{\Omega \max\{M_f^2, M_g^2\}}{\mu\varepsilon} \left(\ln\frac{1}{\sigma} + \ln\log_2 \frac{\mu R_0^2}{2\varepsilon}\right).
\end{aligned}$$

□

8.5 Discussion

We conclude with several remarks concerning possible extensions of the described results.

Obtained results can be easily extended for *composite optimization problems* of the form

$$\min\{f(x) + c(x) : x \in X \subset E, g(x) + c(x) \le 0\}, \tag{8.49}$$

where X is a convex closed subset of finite-dimensional real vector space E, $f : X \to \mathbb{R}$, $g : E \to \mathbb{R}$, $c : X \to \mathbb{R}$ are convex functions. Mirror Descent for unconstrained composite problems was proposed in [11], see also [29] for corresponding version of Dual Averaging [22]. To deal with composite problems (8.49), the Mirror Descent step should be changed to

$$x_+ = \text{Mirr}[x](p) = \arg\min_{u \in X}\left\{\langle p, u\rangle + d(u) + c(u) - \langle \nabla d(x), u\rangle\right\} \quad \forall x \in X^0,$$

where X^0 is defined in Sect. 8.2. The counterpart of Lemma 1 is as follows.

Lemma 6 *Let f be some convex function over a convex closed set X, $h > 0$ be a stepsize, $x \in X^0$. Let the point x_+ be defined by $x_+ = \mathrm{Mirr}[x](h \cdot (\nabla f(x) + \Delta))$, where $\Delta \in E^*$. Then, for any $u \in X$,*

$$\begin{aligned} & h \cdot \big(f(x) - f(u) + c(x_+) - c(u) + \langle \Delta, x - u \rangle\big) \\ & \quad \le h \cdot \langle \nabla f(x) + \Delta, x - u \rangle - h \cdot \langle \nabla c(x_+), u - x_+ \rangle \\ & \quad \le \frac{h^2}{2} \|\nabla f(x) + \Delta\|_{E,*}^2 + V[x](u) - V[x_+](u). \end{aligned}$$

We considered restarting Mirror Descent only in the case of strongly convex functions. A possible extension can be in applying the restart technique to the case of uniformly convex functions f and g introduced in [26] and satisfying

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_E^\rho, \quad x, y \in X,$$

where $\rho \ge 2$, and the same holds for g. Restarting Dual Averaging [22] to obtain subgradient methods for minimizing such functions without functional constraints, both in deterministic and stochastic setting, was suggested in [13]. Another option is, as it was done in [27] for deterministic unconstrained problems, to use sharpness condition of f and g

$$\mu \left(\min_{x_* \in X_*} \|x - x_*\|_E \right)^\rho \le f(x) - f_*, \quad \forall x \in X,$$

where f_* is the minimum value of f, X_* is the set of minimizers of f in Problem (8.1), and the same holds for g.

In stochastic setting, motivated by randomization for deterministic problems, we considered only problems with available values of g. As it was done in [14], one can consider more general problems of minimizing an expectation of a function under inequality constraint given by $\mathbb{E}G(x, \eta) \le 0$, where η is random vector. In this setting one can deal only with stochastic approximation of this inequality constraint.

Acknowledgements The authors are very grateful to Anatoli Juditsky, Arkadi Nemirovski and Yurii Nesterov for fruitful discussions. The research by P. Dvurechensky and A. Gasnikov presented in Section 4 was conducted in IITP RAS and supported by the Russian Science Foundation grant (project 14-50-00150). The research by F. Stonyakin presented in subsection 3.3 was partially supported by the grant of the President of the Russian Federation for young candidates of sciences, project no. MK-176.2017.1.

References

1. A. Anikin, A. Gasnikov, A. Gornov, Randomization and sparsity in huge-scale optimization on an example of mirror descent. Proc. Moscow Inst. Phys. Technol. **8**(1), 11–24 (2016) (in Russian). arXiv:1602.00594

2. A. Bayandina, Adaptive mirror descent for constrained optimization, in *2017 Constructive Nonsmooth Analysis and Related Topics (Dedicated to the Memory of VF Demyanov) (CNSA)*, May 2017, pp. 1–4
3. A. Bayandina, Adaptive stochastic mirror descent for constrained optimization, in *2017 Constructive Nonsmooth Analysis and Related Topics (Dedicated to the Memory of VF Demyanov) (CNSA)*, May 2017, pp. 1–4
4. A. Bayandina, A.G.E. Gasnikova, S. Matsievsky, Primal-dual mirror descent for the stochastic programming problems with functional constraints. Comput. Math. Math. Phys. **58** (2018). arXiv:1604.08194
5. A. Beck, M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003). ISSN: 0167-6377
6. A. Beck, A. Ben-Tal, N. Guttmann-Beck, L. Tetruashvili, The comirror algorithm for solving nonsmooth constrained convex problems. Oper. Res. Lett. **38**(6), 493–498 (2010). ISSN: 0167-6377
7. A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, 2001)
8. A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization (Lecture Notes)*. Personal web-page of A. Nemirovski (2015)
9. A. Birjukov, *Optimization Methods: Optimality Conditions in Extremal Problems*. Moscow Institute of Physics and Technology (2010, in Russian)
10. S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, 2013)
11. J.C. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, Composite objective mirror descent, in *COLT 2010 the 23rd Conference on Learning Theory* (2010), pp. 14–26. ISBN: 9780982252925
12. A. Juditsky, A. Nemirovski, First order methods for non-smooth convex large-scale optimization, I: general purpose methods, in *Optimization for Machine Learning*, ed. by S.J. Wright, S. Sra, S. Nowozin (MIT Press, Cambridge, 2012), pp. 121–184
13. A. Juditsky, Y. Nesterov, Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stochastic Syst. **4**(1), 44–80 (2014)
14. G. Lan, Z. Zhou, Algorithms for stochastic optimization with expectation constraints (2016). arXiv preprint arXiv:1604.03887
15. A. Nedic, S. Lee, On stochastic subgradient mirror-descent algorithm with weighted averaging. SIAM J. Optim. **24**(1), 84–107 (2014)
16. A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
17. A. Nemirovskii, Efficient methods for large-scale convex optimization problems. Ekonomika i Matematicheskie Metody **15** (1979, in Russian)
18. A. Nemirovskii, Y. Nesterov, Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. **25**(2), 21–30 (1985). ISSN: 0041-5553
19. A. Nemirovsky, D. Yudin, *Problem Complexity and Method Efficiency in Optimization* (Wiley, New York, 1983)
20. Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**(2), 372–376 (1983)
21. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course* (Kluwer Academic Publishers, Norwell, 2004)
22. Y. Nesterov, Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009). ISSN: 1436-4646. First appeared in 2005 as CORE discussion paper 2005/67
23. Y. Nesterov, New primal-dual subgradient methods for convex problems with functional constraints, 2015. http://lear.inrialpes.fr/workshop/osl2015/slides/osl2015_yurii.pdf
24. Y. Nesterov, Subgradient methods for convex functions with nonstandard growth properties, 2016. http://www.mathnet.ru:8080/PresentFiles/16179/growthbm_nesterov.pdf
25. B. Polyak, A general method of solving extremum problems. Sov. Math. Dokl. **8**(3), 593–597 (1967)

26. B. Polyak, Existence theorems and convergence of minimizing sequences in extremum problems with restrictions. Sov. Math. Dokl. **7**, 72–75 (1967)
27. V. Roulet, A. d'Aspremont, Sharpness, restart and acceleration (2017). arXiv preprint arXiv:1702.03828
28. N.Z. Shor, Generalized gradient descent with application to block programming. Kibernetika **3**(3), 53–55 (1967)
29. L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**, 2543–2596 (2010). ISSN: 1532-4435