PRIMAL SUBGRADIENT METHODS WITH PREDEFINED STEP SIZES

Yurii Nesterov

REPRINT | 3314

UCLouvain

CORE | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-024-02456-9

Primal Subgradient Methods with Predefined Step Sizes

Yurii Nesterov[1]

Received: 16 November 2023 / Accepted: 5 May 2024

Abstract
In this paper, we suggest a new framework for analyzing primal subgradient methods for nonsmooth convex optimization problems. We show that the classical step-size rules, based on normalization of subgradient, or on knowledge of the optimal value of the objective function, need corrections when they are applied to optimization problems with constraints. Their proper modifications allow a significant acceleration of these schemes when the objective function has favorable properties (smoothness, strong convexity). We show how the new methods can be used for solving optimization problems with functional constraints with a possibility to approximate the optimal Lagrange multipliers. One of our primal-dual methods works also for unbounded feasible set.

Keywords Convex optimization · Nonsmooth optimization · Subgradient methods · Constrained problems · Optimal Lagrange multipliers

1 Introduction

Motivation. The first method for unconstrained minimization of nonsmooth convex function was proposed in [8]. This was a primal subgradient method

$$x_{k+1} = x_k - h_k s_k, \quad s_k = f'(x_k), \quad k \geq 0, \tag{1.1}$$

with constant step sizes $h_k \equiv h > 0$, where $f'(x_k)$ is a subgradient of the objective function at the point x_k. In the next years, there were developed several strategies for

Communicated by Boris S. Mordukhovich.

This paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No 788368). It was also supported by Multidisciplinary Institute in Artificial intelligence MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

✉ Yurii Nesterov
Yurii.Nesterov@uclouvain.be

[1] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Ottignies-Louvain-la-Neuve, Belgium

choosing the steps (see [7] for historical remarks and references). Among them, the most important one is the rule of the *first-order* divergent series:

$$h_k > 0, \quad h_k \to 0, \quad \sum_{k=0}^{\infty} h_k = \infty, \tag{1.2}$$

with the optimal choice $h_k = O(k^{-1/2})$. As a variant, it is possible to use in (2.9) the normalized directions

$$s_k = \frac{f'(x_k)}{\|f'(x_k)\|}. \tag{1.3}$$

Another alternative for the step sizes is based on the known optimal value f^* [7]:

$$h_k = \frac{f(x_k) - f^*}{\|f'(x_k)\|^2}. \tag{1.4}$$

In both cases, the corresponding schemes, as applied to functions with bounded subgradients, have the optimal rate of convergence $O(k^{-1/2})$, established for the best value of the objective function observed during the minimization process [2]. The presence of simple set constraints was treated just by applying to the minimization sequence an Euclidean projection onto the feasible set.

The next important advancement in this area is related to development of the *mirror descent method* (MDM) ([2], see also [1]). In this scheme, the main information is accumulated in the dual space in the form of aggregated subgradients. For defining the next test point, this object is mapped (*mirrored*) to the primal space by a special prox-function, related to a general norm. Thus, we get an important possibility of describing the topology of convex sets by the appropriate norms.

After this discovery, during several decades, the research activity in this field was concentrated on the development of dual schemes. One of the drawbacks of the classical MDM is that the new subgradients are accumulated with the vanishing weights h_k. It was corrected in the framework of *dual averaging* [4], where the aggregation coefficients can be even increasing, and the convergence in the primal space is achieved by applying some vanishing scaling coefficients. Another drawback is related to the fact that the convergence guarantees are traditionally established only for the best values of the objective function. This inconvenience was eliminated by development of *quasi-monotone dual methods* [6], where the rate of convergence is proved for all points of the minimization sequence.

Thus, at some moment, primal methods were almost forgotten. However, in this paper we are going to show that in some situations the primal schemes are very useful. Moreover, there is still space for improvement of the classical methods. Our optimism is supported by the following observations.

Firstly, from the recent developments in Optimization Theory, it becomes clear that the size of subgradients of objective functions for the problems with simple set constraints must be defined differently. Hence, the usual norms in the rules (1.3) and 1.4) can be replaced by more appropriate objects.

Secondly, for an important class of *quasi-convex functions*, linear approximations do not work properly. Hence, for corresponding optimization problems, only the primal schemes can be used. Finally, as we will see, the proper primal schemes provide us with a very simple and natural possibility for approximating the optimal Lagrange multipliers for problems with functional constraints, eliminating somehow the heavy machinery, which is typical for the methods based on Augmented Lagrangian.

Contents. In Sect. 2, we present a new subgradient method for minimizing the quasi-convex function on a simple set. Its justification is based on a new concept of *directional proximity measure*, which is a generalization of the old technique initially presented in [3]. In this method, we apply an *indirect strategy* for choosing the step size, which needs a solution of a simple univariate equation. In unconstrained case, this strategy is reduced to the normalization step (1.3). The main advantage of the new method is the possibility of automatic acceleration for functions with Hölder-continuous gradients.

In Sect. 3, we present a method for solving the composite minimization problem with a max-type objective function. For choosing the step size, we use a proper generalization of the rule (1.4), based on an optimal value of the objective. This method admits a linear rate of convergence for smooth strongly convex functions (see Sect. 4). Note that a simple example demonstrates that the classical rule does not benefit from strong convexity. Method of Sect. 3 automatically accelerates on functions with Hölder continuous gradient.

In Sect. 5, we consider a minimization problem with a single max-type constraint containing an additive composite term. For this problem, we present a switching strategy, where the steps for the objective function are based on the rule of Sect. 2, and for improving the feasibility we use the step-size strategy of Sect. 3. For controlling the step sizes, we suggest a new rule of the *second-order divergent series*:

$$\tau_k \geq \tau_{k+1} > 0, \quad \sum_{k=0}^{\infty} \tau_k^2 = \infty.$$

For the bounded feasible sets, it eliminates an unpleasant logarithmic factor in the convergence rate. The method automatically accelerates for the problem with smooth functional components. It is interesting that the rates of convergence for the objective function and the constraints can be different.

The remaining sections of the paper are devoted to the methods, which can approximate optimal Lagrange multipliers for convex problems with functional inequality constraints. In Sect. 6, we consider the simplest switching strategy of this type, where for the steps with the objective function we use the rule of Sect. 2, and for the steps with violated constraints we use the rule of Sect. 3. In the method of Sect. 7, both steps are based on the rule of Sect. 2. In both cases, we obtain the rates of convergence for infeasibility of the generated points, and the upper bound for the duality gap, computed for the simple estimates of the optimal dual multipliers. Such an estimate is formed as a sum of steps at active iterations for each violated constraint divided by the sum of steps at iterations when the objective function was active.

In Sect. 8, we provide the theoretical guarantees for our estimates of the optimal Lagrange multipliers in terms of the value of dual function. They depend on the depth of Slater condition of our problem. Finally, in Sect. 9, we present a switching method,

which can generate the approximate dual multipliers for problems with unbounded feasible set.

Notation. Denote by $\mathbb{E}$ a finite-dimensional real vector space, and by $\mathbb{E}^*$ its dual space composed by linear functions on $\mathbb{E}$. For such a function $s \in \mathbb{E}^*$, denote by $\langle s, x \rangle$ its value at $x \in \mathbb{E}$. For measuring distances in $\mathbb{E}$, we use an arbitrary norm $\|\cdot\|$. The corresponding dual norm is defined in a standard way:

$$\|g\|_* \;=\; \max_{x \in \mathbb{E}} \left\{ \langle g, x \rangle, \;\; \|x\| \le 1 \right\}, \quad g \in \mathbb{E}^*.$$

Sometimes it is convenient to measure distances in $\mathbb{E}$ by Euclidean norm $\|\cdot\|_B$. It is defined by a self-adjoint positive-definite linear operator $B : \mathbb{E} \to \mathbb{E}^*$ in the following way:

$$\|x\|_B = \langle Bx, x \rangle^{1/2}, \;\; x \in \mathbb{E}, \quad \|g\|_B^* \;=\; \langle g, B^{-1} g \rangle^{1/2}, \;\; g \in \mathbb{E}^*. \tag{1.5}$$

In case of $\mathbb{E} = \mathbb{R}^n$, for $x \in \mathbb{R}^n$, we use the notation $\|x\|_2^2 = \sum\limits_{i=1}^{n} (x^{(i)})^2$.

For a differentiable function $f(\cdot)$ with convex and open domain $\mathrm{dom}\, f \subseteq \mathbb{E}$, denote by $\nabla f(x) \in \mathbb{E}^*$ its gradient at point $x \in \mathrm{dom}\, f$. If f is convex, it can be used for defining the *Bregman distance* between two points $x, y \in \mathrm{dom}\, f$:

$$\beta_f(x, y) \;=\; f(y) - f(x) - \langle \nabla f(x), y - x \rangle \;\ge\; 0. \tag{1.6}$$

In this paper, we develop new *proximal-gradient methods* based on a predefined *prox-function* $d(\cdot)$, which can be restricted to a convex open domain $\mathrm{dom}\, d \subseteq \mathbb{E}$. This domain always contains the basic feasible set of the corresponding optimization problem. We assume that $d(\cdot)$ is continuously differentiable and *strongly convex* on $\mathrm{dom}\, d$ with parameter one:

$$d(y) \ge d(x) + \langle \nabla d(x), y - x \rangle + \;\tfrac{1}{2}\|y - x\|^2, \quad x, y \in \mathrm{dom}\, d. \tag{1.7}$$

Thus, combining the definition (1.6) with inequality (1.7), we get

$$\beta_d(x, y) \ge \;\tfrac{1}{2}\|x - y\|^2, \quad x, y \in \mathrm{dom}\, d. \tag{1.8}$$

2 Subgradient method for quasi-convex problems

Consider the following constrained optimization problem:

$$\min_{x \in Q} \; f_0(x), \tag{2.1}$$

where the function $f_0(\cdot)$ is closed and quasi-convex on $\mathrm{dom}\, f_0$ and the set $Q \subseteq \mathrm{dom}\, f_0$ is closed and convex. Denote by x^* one of the optimal solutions of (2.1) and let $f_0^* = f_0(x^*)$. We assume that

$$x^* \in \mathrm{int}\, \mathrm{dom}\, f_0. \tag{2.2}$$

Let us assume that at any point $x \in Q$ it is possible to compute a vector $f_0'(x) \in \mathbb{E}^* \setminus \{0\}$, satisfying the following condition: for any $y \in \mathbb{E}$, we have:

$$\langle f_0'(x), y - x \rangle \geq 0 \Rightarrow f_0(y) \ \geq \ f_0(x). \tag{2.3}$$

(If $y \notin \operatorname{dom} f_0$, then $f_0(y) \stackrel{\text{def}}{=} +\infty$.) If $f_0(\cdot)$ is differentiable at x, then $f_0'(x) = \nabla f_0(x)$.

In order to justify the rate of convergence of our scheme, we need the following characteristic of problem (2.1):

$$\mu(r) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{E}} \left\{ f_0(x) - f_0^* : \ \|x - x^*\| \leq r \right\}, \quad r \geq 0. \tag{2.4}$$

If $y \notin \operatorname{dom} f_0$ and $r = \|y - x^*\|$, then $\mu(r) = +\infty$. In view of assumption (2.2), this function is finite at least in some neighborhood of the origin.

We say that vector $s \in \mathbb{E}$ defines a *direction* in $\mathbb{E}$ if $\|s\| = 1$. If $\langle f_0'(x), s \rangle > 0$, we call it a *recession direction* of function $f_0(\cdot)$ at point $x \in Q$. Using such a direction, we can define the *directional proximity measure* of point $x \in Q$ as follows:

$$\delta_s(x) \stackrel{\text{def}}{=} \frac{\langle f_0'(x), x - x^* \rangle}{\langle f_0'(x), s \rangle} \ \geq \ 0. \tag{2.5}$$

Lemma 1 *Let s be a recession direction at point $x \in Q$. Then*

$$f_0(x) - f_0^* \leq \mu(\delta_s(x)). \tag{2.6}$$

Proof Indeed, let us define $y = x^* + \delta_s(x) s$. Then

$$\langle f_0'(x), y \rangle = \langle f_0'(x), x^* \rangle + \delta_s(x) \langle f_0'(x), s \rangle \stackrel{(2.5)}{=} \langle f_0'(x), x \rangle.$$

Since $f_0(\cdot)$ is quasi-convex, this means that $f_0(y) \stackrel{(2.3)}{\geq} f_0(x)$. Therefore,

$$f_0(x) - f_0^* \leq f_0(y) - f_0^* \stackrel{(2.4)}{\leq} \mu(\|y - x^*\|) \ = \ \mu(\delta_s(x)).$$

□

In our analysis, we use the following univariate functions:

$$\varphi_{\bar{x}}(\lambda) = \max_{x \in Q} \left\{ \lambda \langle f_0'(\bar{x}), \bar{x} - x \rangle - \beta_d(\bar{x}, x) \right\}, \quad \lambda \geq 0, \tag{2.7}$$

where $\bar{x} \in Q$. The optimal solution of this optimization problem is denoted by $T_{\bar{x}}(\lambda)$. Note that function $\varphi_{\bar{x}}(\cdot)$ is convex and continuously differentiable with the following derivative:

$$\varphi_{\bar{x}}'(\lambda) = \langle f_0'(\bar{x}), \bar{x} - T_{\bar{x}}(\lambda) \rangle. \tag{2.8}$$

Thus, $\varphi_{\bar{x}}(0) = 0$, $T_{\bar{x}}(0) = \bar{x}$, and $\varphi'_{\bar{x}}(0) = 0$. Consequently, $\varphi_{\bar{x}}(\lambda) \geq 0$ for all $\lambda \geq 0$.

The basic subgradient method for solving the problem (2.1) looks as follows.

Basic Subgradient Method

Initialization. Choose $x_0 \in Q$.

k**th iteration** ($k \geq 0$)**.** a) Choose step-size parameter $\lambda_k > 0$.

b) Compute $x_{k+1} = T_{x_k}(\lambda_k) \stackrel{\text{def}}{=} \arg\min_{x \in Q} \left\{ \lambda_k \langle f_0'(x_k), x \rangle + \beta_d(x_k, x) \right\}$. (2.9)

Lemma 2 *At each iteration $k \geq 0$ of method (2.9), for any $x \in Q$, we have*

$$\beta_d(x_{k+1}, x) \leq \beta_d(x_k, x) - \lambda_k \langle f_0'(x_k), x_k - x \rangle + \varphi_{x_k}(\lambda_k). \tag{2.10}$$

Proof Note that the first-order optimality condition for the problem at Step b) can be written as follows:

$$\langle \lambda_k f_0'(x_k) + \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle \geq 0, \quad \forall x \in Q. \tag{2.11}$$

Therefore, we have

$$\begin{aligned}
\beta_d(x_{k+1}, x) &\stackrel{(1.6)}{=} d(x) - d(x_{k+1}) - \langle \nabla d(x_{k+1}), x - x_{k+1} \rangle \\
&= \beta_d(x_k, x) + d(x_k) + \langle \nabla d(x_k), x - x_k \rangle - d(x_{k+1}) - \langle \nabla d(x_{k+1}), x - x_{k+1} \rangle \\
&\stackrel{(2.11)}{\leq} \beta_d(x_k, x) + d(x_k) + \langle \nabla d(x_k), x - x_k \rangle - d(x_{k+1}) \\
&\quad + \lambda_k \langle f_0'(x_k), x - x_{k+1} \rangle - \langle \nabla d(x_k), x - x_{k+1} \rangle \\
&= \beta_d(x_k, x) + \lambda_k \langle f_0'(x_k), x - x_{k+1} \rangle - \beta_d(x_k, x_{k+1}) \\
&= \beta_d(x_k, x) - \lambda_k \langle f_0'(x_k), x_k - x \rangle + \varphi_{x_k}(\lambda_k).
\end{aligned}$$

□

In our method, we will use an *indirect control* of the *dual* step sizes $\{\lambda_k\}_{k \geq 0}$, which is based on a predefined sequence of *primal* step-size parameters $\{h_k\}_{k \geq 0}$. As we will prove by Lemma 3, the convergence of the process can be derived, for example, from the following standard conditions:

$$\{h_k\}_{k \geq 0}: \quad \text{a) } h_k > 0, \quad \text{b) } \sum_{k=0}^{\infty} h_k = \infty. \tag{2.12}$$

Then, inequality (2.10) justifies the choice of the *dual* step-size parameter λ_k as a solution to the following equation:

$$\boxed{\varphi_{x_k}(\lambda_k) = \tfrac{1}{2}h_k^2, \quad k \geq 0} \tag{2.13}$$

Example 1 If $Q = \mathbb{E}$ and $d(x) = \frac{1}{2}\|x\|_B^2$, then $\varphi_{\bar{x}}(\lambda) \overset{(2.7)}{=} \frac{\lambda^2}{2}(\|f_0'(\bar{x})\|_B^*)^2$, and equation (2.13) gives us $\lambda_k = h_k / \|f_0'(x_k)\|_B^*$. In this case, method (2.9) coincides with the classical variant of subgradient method $x_{k+1} = x_k - h_k \frac{f_0'(x_k)}{\|f_0'(x_k)\|_B^*}$. However, for $Q \neq \mathbb{E}$, the rule (2.13) allows a proper scaling of the step size by the boundary of feasible set. □

At each iteration of method (2.9), the corresponding value of λ_k can be found by an efficient one-dimensional search procedure, based, for example, on a Newton-type scheme. Since the latter scheme has local quadratic convergence, we make a plausible assumption that it is possible to compute an exact solution of the equation (2.13). This solution has the following important interpretation (in view of (2.8)):

$$\lambda_k = \max_{\lambda \geq 0} \left\{ \lambda : \ \varphi_{x_k}(\lambda) \leq \tfrac{1}{2}h_k^2 \right\}. \tag{2.14}$$

At the same time, we have $\langle \lambda_k f_0'(x_k) + \nabla d(x_{k+1}) - \nabla d(x_k), x_k - x_{k+1} \rangle \overset{(2.11)}{\geq} 0$. Thus

$$\begin{aligned} \lambda_k \langle f_0'(x_k), x_k - x_{k+1} \rangle &\overset{(1.6)}{\geq} \beta_d(x_k, x_{k+1}) + \beta_d(x_{k+1}, x_k) \\ &\overset{(1.8)}{\geq} \beta_d(x_k, x_{k+1}) + \tfrac{1}{2}\|x_k - x_{k+1}\|^2. \end{aligned}$$

Hence

$$\tfrac{1}{2}h_k^2 \overset{(2.13)}{=} \varphi_{x_k}(\lambda_k) \overset{(2.7)}{\geq} \tfrac{1}{2}\|x_k - x_{k+1}\|^2, \quad k \geq 0. \tag{2.15}$$

Our complexity bounds follow from the rate of convergence of the following values:

$$\delta_k \overset{\text{def}}{=} \delta_{s_k}(x_k), \quad s_k \overset{\text{def}}{=} \tfrac{x_k - x_{k+1}}{\|x_k - x_{k+1}\|}.$$

Lemma 3 *Let condition (2.13) be satisfied. Then for any $N \geq 0$, we have*

$$\sum_{k=0}^{N} h_k \delta_k \leq \beta_d(x_0, x^*) + \tfrac{1}{2}\sum_{k=0}^{N} h_k^2. \tag{2.16}$$

Proof Indeed, in view of inequality (2.10), for $r_k = \beta_d(x_k, x^*)$, we have

$$r_{k+1} \leq r_k - \lambda_k \langle f_0'(x_k), x_k - x^* \rangle + \varphi_{x_k}(\lambda_k) \overset{(2.13)}{=} r_k - \lambda_k \langle f_0'(x_k), x_k - x^* \rangle + \tfrac{1}{2}h_k^2.$$

On the other hand,

$$\lambda_k \langle f_0'(x_k), x_k - x_{k+1} \rangle \overset{(2.13)}{=} \beta_d(x_k, x_{k+1}) + \tfrac{1}{2}h_k^2 \overset{(1.8)}{\geq} \tfrac{1}{2}\|x_k - x_{k+1}\|^2 + \tfrac{1}{2}h_k^2 \geq h_k \|x_k - x_{k+1}\|.$$

Hence,

$$r_{k+1} \leq r_k - h_k \delta_k + \tfrac{1}{2}h_k^2. \tag{2.17}$$

Summing up these inequalities for $k = 0, \ldots, N$, we obtain inequality (2.16). □

Let us look now at one example of the rate of convergence of method (2.9) for an objective function from a nonstandard problem class. For simplicity, let us measure distances by Euclidean norm $\|x\|_B$ (see (1.5)). In this case, we can take $d(x) = \frac{1}{2}\|x\|_B^2$ and get $\beta_d(x, y) = \frac{1}{2}\|x - y\|_B^2$. Let us assume that the function $f_0(\cdot)$ in problem (2.1) is p-times continuously differentiable on $\mathbb{E}$ and its pth derivative is Lipschitz-continuous with constant L_p. Then we can bound the function $\mu(\cdot)$ in (2.4) as follows:

$$\mu(r) \leq \sum_{i=1}^{p} \frac{r^i}{i!} \|D^i f_0(x^*)\| + \frac{L_p}{(p+1)!} r^{p+1}, \quad r \geq 0, \tag{2.18}$$

where all norms for derivatives are induced by $\|\cdot\|_B$.

Let us fix the total number of steps $N \geq 1$ and assume the bound $R_0 \geq \|x_0 - x^*\|_B$ be available. Then, defining the step sizes

$$h_k = h \overset{\text{def}}{=} \frac{R_0}{\sqrt{N+1}}, \quad 0 \leq k \leq N, \tag{2.19}$$

we get $\delta_N^* = \min\limits_{0 \leq k \leq N} \delta_k \overset{(2.16)}{\leq} \frac{R_0}{\sqrt{N+1}}$. Hence, in view of inequality (2.18), we have

$$f_N^* \overset{\text{def}}{=} \min_{0 \leq k \leq N} f(x_k) \leq f_0^* + \sum_{i=1}^{p} \frac{1}{i!} \|D^i f_0(x^*)\| \left(\frac{R_0}{\sqrt{N+1}}\right)^i + \frac{L_p}{(p+1)!} \left(\frac{R_0}{\sqrt{N+1}}\right)^{p+1}. \tag{2.20}$$

Note that the first p coefficients in estimate (2.20) depend on the local properties of the objective function at the solution, and only the last term employs the global Lipschitz constant for the pth derivative. Clearly, we do not need to know the bounds for all these derivatives in order to define the step-size strategy (2.19).

3 Step-size control for max-type convex problems

Let us consider now the following problem of *composite optimization*

$$F_* \stackrel{\text{def}}{=} \min_{x \in \text{dom}\, \psi} \Big\{ F(x) = f(x) + \psi(x) \Big\}, \tag{3.1}$$

where $\psi(\cdot)$ is a simple closed convex function, and

$$f(x) = \max_{1 \le i \le m} f_i(x), \quad x \in \text{dom}\, \psi, \tag{3.2}$$

with all $f_i(\cdot)$, $1 \le i \le m$, being closed and convex on $\text{dom}\, \psi$. Denote by

$$\ell_{\bar{x}}(x) = \max_{1 \le i \le m} [\, f_i(\bar{x}) + \langle f_i'(\bar{x}), x - \bar{x} \rangle \,] \le f(x), \quad x \in \text{dom}\, \psi,$$

the linearization of function $f(\cdot)$, and by $x_* \in \text{dom}\, \psi$ an optimal solution of this problem.

Similarly to (2.7), let us define the following univariate functions:

$$\hat{\varphi}_{\bar{x}}(\lambda) = \max_{x \in \text{dom}\, \psi} \Big\{ \lambda \Big[F(\bar{x}) - \ell_{\bar{x}}(x) - \psi(x) \Big] - \beta_d(\bar{x}, x) \Big\}, \quad \lambda \ge 0, \tag{3.3}$$

where $\bar{x} \in \text{dom}\, \psi$. The unique optimal solution of this optimization problem is denoted by $\hat{T}_{\bar{x}}(\lambda)$. Note that function $\hat{\varphi}_{\bar{x}}(\cdot)$ is convex and continuously differentiable with the following derivative:

$$\hat{\varphi}_{\bar{x}}'(\lambda) = -\Big[\ell_{\bar{x}}(\hat{T}) + \psi(\hat{T}) - F(\bar{x}) \Big], \quad \lambda \ge 0, \tag{3.4}$$

where $\hat{T} = \hat{T}_{\bar{x}}(\lambda)$. Thus, $\hat{\varphi}_{\bar{x}}(0) = 0$, $\hat{T}_{\bar{x}}(0) = \bar{x}$, and $\hat{\varphi}_{\bar{x}}'(0) = 0$. Hence, $\hat{\varphi}_{\bar{x}}(\lambda) \ge 0$ for all $\lambda \ge 0$. Let us prove the following variant of Lemma 2.

Lemma 4 *Let $\bar{x} \in \text{dom}\, \psi$ and $\hat{T} = \hat{T}_{\bar{x}}(\lambda)$ for some $\lambda \ge 0$. Then*

$$\beta_d(\hat{T}, x_*) \le \beta_d(\bar{x}, x_*) + \lambda \Big[\ell_{\bar{x}}(x_*) + \psi(x_*) - F(\bar{x}) \Big] + \hat{\varphi}_{\bar{x}}(\lambda). \tag{3.5}$$

Proof In view of the first-order optimality condition for the minimization problem in (3.3), for all $x \in \text{dom}\, \psi$, we have

$$\langle \nabla d(\hat{T}) - \nabla d(\bar{x}), x - \hat{T} \rangle + \lambda \Big[\ell_{\bar{x}}(x) + \psi(x) \Big] \ge \lambda \Big[\ell_{\bar{x}}(\hat{T}) + \psi(\hat{T}) \Big]. \tag{3.6}$$

Therefore, we get

$$\begin{aligned}\beta_d(\hat{T}, x_*) &= \beta_d(\bar{x}, x_*) + d(\bar{x}) + \langle \nabla d(\bar{x}), x_* - \bar{x}\rangle - d(\hat{T}) - \langle \nabla d(\hat{T}), x^* - \hat{T}\rangle \\ &\overset{(3.6)}{\leq} \beta_d(\bar{x}, x_*) + d(\bar{x}) + \langle \nabla d(\bar{x}), x_* - \bar{x}\rangle - d(\hat{T}) \\ &\quad -\langle \nabla d(\bar{x}), x_* - \hat{T}\rangle + \lambda\Big[\ell_{\bar{x}}(x_*) + \psi(x_*) - \ell_{\bar{x}}(\hat{T}) - \psi(\hat{T})\Big] \\ &\overset{(3.3)}{=} \beta_d(\bar{x}, x_*) + \lambda\Big[\ell_{\bar{x}}(x_*) + \psi(x_*) - F(\bar{x})\Big] + \hat{\varphi}_{\bar{x}}(\lambda).\end{aligned}$$

□

In this section, we analyze the following optimization scheme.

Basic Composite Subgradient Method

Initialization. Choose $x_0 \in \operatorname{dom} \psi$.

kth iteration ($k \geq 0$).

a) Choose step-size parameter $\lambda_k > 0$.

b) Compute $x_{k+1} = \hat{T}_{x_k}(\lambda_k)$. (3.7)

Suppose that the optimal value F_* is known. In view of convexity of the function $f(\cdot)$, we have

$$\ell_{\bar{x}}(x_*) + \psi(x_*) - F(\bar{x}) \leq F_* - F(\bar{x}).$$

Hence, for points $\{x_k\}_{k\geq 0}$, generated by method (3.7), we have

$$\beta_d(x_{k+1}, x_*) \overset{(3.5)}{\leq} \beta_d(x_k, x_*) + \Big(\hat{\varphi}_{x_k}(\lambda_k) - \lambda_k[F(x_k) - F_*]\Big). \tag{3.8}$$

This observation explains the following step-size strategy:

$$\boxed{\lambda_k = \arg\min_{\lambda\geq 0}\Big\{\ \hat{\varphi}_{x_k}(\lambda) - \lambda[F(x_k) - F_*]\ \Big\}} \tag{3.9}$$

This strategy has a natural optimization interpretation.

Lemma 5 *Let λ_k be defined by (3.9). Then*

$$\hat{T}_{x_k}(\lambda_k) = \arg \min_{x \in \mathrm{dom}\, \psi} \Big\{ \beta_d(x_k, x) : \ \ell_{x_k}(x) + \psi(x) \le F_* \Big\}. \tag{3.10}$$

Proof Indeed,

$$\begin{aligned}
&\min_{\lambda \ge 0} \Big\{ \hat{\varphi}_{x_k}(\lambda) - \lambda [F(x_k) - F_*] \Big\} \\
&\overset{(3.3)}{=} \min_{\lambda \ge 0} \max_{x \in \mathrm{dom}\, \psi} \Big\{ -\lambda \Big[\ell_{x_k}(x) + \psi(x) - F(x_k) \Big] - \beta_d(x_k, x) - \lambda [F(x_k) - F_*] \Big\} \\
&= \max_{x \in \mathrm{dom}\, \psi} \min_{\lambda \ge 0} \Big\{ -\beta_d(x_k, x) + \lambda \Big[F_* - \ell_{x_k}(x) - \psi(x) \Big] \Big\} \\
&= - \min_{x \in \mathrm{dom}\, \psi} \Big\{ \beta_d(x_k, x) : \ \ell_{x_k}(x) + \psi(x) \le F_* \Big\}.
\end{aligned}$$

□

Lemma 5 has two important consequences allowing us to estimate the rate of convergence of method (3.7) with the step-size rule (3.9). Namely, for any $k \ge 0$ we have

$$\beta_d(x_{k+1}, x_*) \overset{(3.8)}{\le} \beta_d(x_k, x_*) - \beta_d(x_k, x_{k+1}), \tag{3.11}$$

$$\ell_{x_k}(x_{k+1}) + \psi(x_{k+1}) \overset{(3.10)}{\le} F_*. \tag{3.12}$$

Note that method (3.7) is not monotone. However, inequality (3.11) gives us a global rate of convergence for the following characteristic:

$$\rho_k \overset{\mathrm{def}}{=} \min_{0 \le i \le k} \|x_i - x_{i+1}\| \overset{(1.8)}{\le} \min_{0 \le i \le k} \sqrt{2\beta_d(x_i, x_{i+1})} \overset{(3.11)}{\le} \frac{r_0}{\sqrt{k+1}}, \quad k \ge 0, \tag{3.13}$$

where $r_0 = \beta_d(x_0, x_*)$.

Theorem 1 *Let all functions $f_i(\cdot)$ have Hölder-continuous gradients:*

$$f_i(y) \le f_i(x) + \langle f_i'(x), y - x \rangle + \frac{L_\nu}{1+\nu} \|y - x\|^{1+\nu}, \quad x, y \in \mathrm{dom}\, \psi, \ 1 \le i \le p, \tag{3.14}$$

where $\nu \in [0, 1]$ and $L_\nu \ge 0$. Then for the step-size rule (3.9) in method (3.7) we have

$$F_k \overset{\mathrm{def}}{=} \min_{0 \le i \le k} F(x_i) \ \le \ F_* + \frac{L_\nu r_0^{1+\nu}}{1+\nu} \Big(\frac{1}{k} \Big)^{\frac{1+\nu}{2}}, \quad k \ge 1. \tag{3.15}$$

Proof Indeed, let $\rho_k = \|x_{i_k} - x_{i_k+1}\|$ for some i_k, $0 \le i_k \le k$. Then for any $k \ge 0$ we have

$$F_{k+1} \le F(x_{i_k+1}) \overset{(3.14)}{\le} \ell_{x_{i_k}}(x_{i_k+1}) + \tfrac{L_\nu}{1+\nu}\rho_k^{1+\nu} + \psi(x_{i_k+1})$$

$$\overset{(3.12)}{\le} F_* + \tfrac{L_\nu}{1+\nu}\rho_k^{1+\nu}.$$

It remains to use inequality (3.13). □

Note that the step-size strategy (3.9) does not depend on the Hölder parameter ν in the condition (3.14). Hence, the number of iterations of method (3.7), (3.9), ensuring an ϵ-accuracy in function value, is bounded from above by the following quantity:

$$\inf_{0 \le \nu \le 1} \left[\tfrac{L_\nu}{(1+\nu)\epsilon}\right]^{\frac{2}{1+\nu}} r_0^2. \tag{3.16}$$

To conclude the section, let us show that the step-size rule (3.9) can behave much better than the classical one.

Example 2 Let $\mathbb{E} = \mathbb{R}^2$, $\|x\|^2 = x^T x$, and $f(x) = \frac{1}{2}(x^{(1)})^2 + \frac{1}{2}(x^{(2)} - 1)^2$ with $\nabla f(x) = (x^{(1)}, x^{(2)} - 1)^T$. Define $\psi(x) = \mathrm{Ind}\, Q$, the indicator function of $Q = \left\{x \in \mathbb{R}^2 : x^{(2)} \le 0\right\}$. Then $f_* = 0.5$. Consider the point $x_k = (x_k^{(1)}, 0)^T$. By the classical step-size rule, we have

$$x_{k+1} = \pi_Q\left[x_k - \tfrac{f(x_k) - f_*}{\|\nabla f(x_k)\|^2}\nabla f(x_k)\right] = \pi_Q\left[\begin{pmatrix} x_k^{(1)} \\ 0 \end{pmatrix} - \frac{\left(x_k^{(1)}\right)^2}{2\left(1+\left(x_k^{(1)}\right)^2\right)}\begin{pmatrix} x_k^{(1)} \\ -1 \end{pmatrix}\right].$$

Thus, $x_{k+1}^{(2)} = 0$ and $x_{k+1}^{(1)} = x_k^{(1)} - \dfrac{\left(x_k^{(1)}\right)^3}{2\left(1+\left(x_k^{(1)}\right)^2\right)}$. This means that $x_k^{(1)} = O(k^{-1/2})$.

On the other hand, the rule (3.10) as applied to the same point x_k defines $x = x_{k+1}$ as an intersection of two lines

$$\left\{x \in \mathbb{R}^2 : \langle \nabla f(x_k), x_k - x\rangle = f(x_k) - f_*\right\}, \quad \left\{x \in \mathbb{R}^2 : x^{(2)} = 0\right\}.$$

This means that $x_k^{(1)}\left(x_k^{(1)} - x_{k+1}^{(1)}\right) = \frac{1}{2}\left(x_k^{(1)}\right)^2$, and we get the linear rate of convergence to the optimal point $x_* = 0$. □

4 Step-size control for problems with smooth strongly convex components

A linear rate of convergence, demonstrated by method (3.7) in the Example 2, provides us with motivation to look at behavior of this method on smooth and strongly convex

problems. Let us introduce the Euclidean metric by (1.5) and define $d(x) = \frac{1}{2}\|x\|_B^2$, $x \in \mathbb{E}$. In this case,

$$\beta_d(x, y) = \frac{1}{2}\|x - y\|_B^2, \quad x, y \in \mathbb{E}.$$

Suppose that all functions $f_i(\cdot)$, $i = 1, \ldots, m$ in (3.2) are continuously differentiable and have Lipschitz-continuous gradients with the same constant $L_f \geq 0$:

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{1}{2} L_f \|x - y\|_B^2, \quad x, y \in \operatorname{dom} \psi. \tag{4.1}$$

In the notation of Sect. 3, these inequalities imply the following upper bound for the objective function of problem (3.1):

$$F(y) \leq \ell_x(y) + \psi(y) + \frac{1}{2} L_f \|x - y\|_B^2, \quad x, y \in \operatorname{dom} \psi. \tag{4.2}$$

Hence, for the sequence of points $\{x_k\}_{k \geq 0}$, generated by method (3.7), we have

$$\begin{aligned} F(x_{k+1}) &\leq \ell_{x_k}(x_{k+1}) + \psi(x_{k+1}) + \frac{1}{2} L_f \|x_k - x_{k+1}\|_B^2 \\ &\overset{(3.12)}{\leq} F_* + \frac{1}{2} L_f \|x_k - x_{k+1}\|_B^2. \end{aligned} \tag{4.3}$$

Thus, we can prove the following statement.

Theorem 2 *Under condition (4.1), the rate of convergence of method (3.7) can be estimated as follows:*

$$\frac{1}{T} \sum_{k=1}^{T} F(x_k) - F_* \leq \frac{1}{2T} L_f \|x_0 - x_*\|_B^2, \quad T \geq 1. \tag{4.4}$$

If in addition functions $f_i(\cdot)$, $i = 1, \ldots, m$, are strongly convex:

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{1}{2} \mu_f \|x - y\|_B^2, \quad x, y \in \operatorname{dom} \psi, \tag{4.5}$$

with $\mu_f > 0$, then the rate of convergence is linear:

$$\|x_k - x_*\|_B^2 \leq \left(\frac{L_f}{\mu_f + L_f} \right)^k \|x_0 - x_*\|_B^2, \quad k \geq 0. \tag{4.6}$$

Proof Indeed, substituting inequality (4.3) into relation (3.11), we have

$$\frac{1}{2}\|x_{k+1} - x_*\|_B^2 \leq \frac{1}{2}\|x_k - x_*\|_B^2 - \frac{1}{L_f}[F(x_{k+1}) - F_*], \quad k \geq 0. \tag{4.7}$$

Summing up these inequalities for $k = 0, \ldots, T - 1$, we get inequality (4.4).

If in addition, the conditions (4.5) are satisfied, then

$$F(y) \geq \ell_x(y) + \psi(y) + \frac{1}{2} \mu_f \|x - y\|_B^2, \quad x, y \in \operatorname{dom} \psi. \tag{4.8}$$

Note that the first-order optimality conditions for problem (3.1) can be written in the following form:

$$\ell_{x_*}(y) + \psi(y) \geq F_*, \quad \forall y \in \operatorname{dom} \psi.$$

Therefore, inequality (4.8) implies that

$$F(y) - F_* \geq \tfrac{1}{2}\mu_f \|y - x_*\|_B^2, \quad \forall y \in \operatorname{dom} \psi.$$

Thus, $\|x_{k+1} - x_*\|_B^2 \overset{(4.7)}{\leq} \frac{L_f}{\mu_f + L_f}\|x_k - x_*\|_B^2$, and we get inequality (4.6). □

5 Convex minimization with max-type composite constraint

Let us show that both step-size strategies described in Sects. 2 and 3 can be unified in one scheme for solving constrained optimization problems. In this section, we deal with the problem in the following *semi-composite* form:

$$\min_{x \in Q} \left\{ f_0(x) : \; F(x) \overset{\text{def}}{=} f(x) + \psi(x) \leq 0 \right\} \tag{5.1}$$

where $\psi(\cdot)$ is a simple closed convex function, $Q \subseteq \operatorname{dom} \psi$ is a closed convex set, and

$$f(x) = \max_{1 \leq i \leq m} f_i(x),$$

with all functions $f_i(\cdot), i = 0, \ldots, m$, being closed and convex on $\operatorname{dom} \psi$. We assume Q to be bounded:

$$\beta_d(x, y) < D, \quad \forall x, y \in Q. \tag{5.2}$$

In order to solve problem (5.1), we propose a method, which combines two different types of iterations. One of them improves the *feasibility* of the current point, and the second one improves its *optimality*.

Iteration of the first type is based on the machinery developed in Sect. 3 with the particular value $F_* = 0$. It is applied to some point $x_k \in Q$.

$$\boxed{\text{Compute } \lambda_k = \arg\min_{\lambda \geq 0} \left\{ \hat{\varphi}_{x_k}(\lambda) - \lambda F(x_k) \right\}. \text{ Set } x_{k+1} \overset{(3.10)}{=} \hat{T}_{x_k}(\lambda_k).} \tag{5.3}$$

Note that for $F(x_k) \leq 0$, we have $\lambda_k = 0$ and $\hat{T}_{x_k}(\lambda_k) = x_k$.

For iteration k of the second type, we need to choose a primal step-size bound $h_k > 0$. Then, at the test point $y_k \in Q$, we define the function $\varphi_{y_k}(\cdot)$ by (2.7) and apply the following rule:

Find λ_k from equation $\varphi_{y_k}(\lambda_k) = \frac{1}{2}h_k^2$. Set $x_{k+1} = T_{y_k}(\lambda_k)$.	(5.4)

Since in both rules parameters λ_k are functions of the test points, we will use shorter notations $T(y_k)$ and $\hat{T}(x_k)$. Consider the following optimization scheme.

Double-Step Subgradient Method for Semi-Composite Problem
Initialization. Choose $x_0 \in \operatorname{dom}\psi$ and sequence of steps $\mathcal{H} = \{h_k\}_{k\geq 0}$.
k**th iteration** ($k \geq 0$)**. 1.** Compute $y_k = \hat{T}(x_k)$.
2. If $\beta_d(x_k, y_k) \geq \frac{1}{2}h_k^2$, then a) $x_{k+1} = y_k$. Else, b) compute $x_{k+1} = T(y_k)$.

(5.5)

Thus, method (5.5) is defined by a sequence of primal step bounds $\mathcal{H} = \{h_k\}_{k\geq 0}$. However, since in (5.5) we apply a switching strategy, it is impossible to say in advance what will be the type of a particular kth iteration. Therefore, as compared with the classical conditions (2.12), we need additional regularity assumptions on $\mathcal{H}$.

It will be convenient to relate this sequence with another sequence of *scaling coefficients* $\mathcal{T} = \{\tau_k\}_{k\geq 0}$, satisfying the *second-order divergence condition* (compare with (2.12)).

Second-Order Divergence Condition
a) $\tau_k \geq \tau_{k+1} > 0$ for any $k \geq 0$. **b)** $\sum\limits_{k=0}^{\infty} \tau_k^2 = +\infty$.

(5.6)

Note that condition (5.6) ensures $\sum\limits_{k=0}^{\infty} \tau_k = +\infty$. Thus, it is stronger than (2.12). In order to transform the sequence $\mathcal{T}$ into *convergence rate* of some optimization process, we need to introduce the following characteristic.

Definition 1 For a sequence $\mathcal{T}$, the integer-valued function $a(k)$, $k \geq 0$, is called the *divergence delay* (of degree two) if $a(k) \geq 0$ is the minimal integer value such that

$$\sum_{i=k}^{k+a(k)} \tau_i^2 \geq 1, \quad k \geq 0. \tag{5.7}$$

Clearly, condition $(5.6)_b$ ensures that all values $a(k)$, $k \geq 0$, are well defined. At the same time, from $(5.6)_a$, we have $a(k+1) \geq a(k)$ for any $k \geq 0$.

Let us give two important examples of such sequences.

Example 3 **(a)** Let us fix an integer $N \geq 0$ and define

$$\tau_k = \frac{1}{\sqrt{N+1}}, \quad k \geq 0. \tag{5.8}$$

Then condition $(5.6)_a$ is valid and $\sum_{i=k}^{k+a(k)} \tau_i^2 = \frac{a(k)+1}{N+1}$. Thus, $a(k) = N$.

(b) Consider the following sequence:

$$\tau_k = \sqrt{\frac{2}{k+1}}, \quad k \geq 0. \tag{5.9}$$

For $k \geq 1$, denote by $S_k = \sum_{i=0}^{k-1} \tau_i^2$. Then $S_1 = 2$, $S_2 = 3$, and the difference

$$S_{2k} - S_k = \sum_{i=k}^{2k-1} \tau_i^2$$

is monotonically increasing in $k \geq 1$. Thus, $a(k) \leq k-1$ for all $k \geq 1$. □

Let us analyze the performance of the method (5.5) with an appropriately chosen sequence $\mathcal{H}$. Namely, let us choose

$$h_k = \sqrt{2D}\, \tau_k, \quad k \geq 0, \tag{5.10}$$

where the sequence $\mathcal{T}$ satisfies condition (5.6) and D is taken from (5.2).

We are interested only in the values of objective function $f_0(\cdot)$ computed at the points with small values of the functional constraint $F(\cdot)$. As we will see, these points are involved in Step 2b). For the total number of steps $N \geq 1 + a(0)$, denote

$$\begin{aligned} k(N) &= \max \left\{ k \geq 0 : \; k + a(k) \leq N - 1 \right\}, \\ \mathcal{F}_N &= \left\{ k : \; k(N) \leq k \leq N-1, \; y_k \xrightarrow{2b)} x_{k+1} \right\}. \end{aligned} \tag{5.11}$$

Let us define the following directional proximity measures:

$$\delta_k = \delta_{s_k}(y_k) \text{ with } s_k = \frac{y_k - x_{k+1}}{\|y_k - x_{k+1}\|}, \; k \in \mathcal{F}_N.$$

We are interested in the rate of convergence to zero of the following characteristic:

$$\delta_N^* = \min_{k \in \mathcal{F}_N} \delta_k.$$

Theorem 3 *For any $N \geq 1 + a(0)$, the number $k(N) \geq 0$ is well defined and $\delta_N^* < h_{k(N)}$.*
Moreover, if all functions $f_i(\cdot)$, $i = 1, \ldots, m$, have Hölder-continuous gradients on Q with parameter $\nu \in [0, 1]$ and constant $L_\nu > 0$, then, for any $k \in \mathcal{F}_N$, we have

$$F(y_k) \leq \tfrac{L_\nu}{1+\nu} h_{k(N)}^{1+\nu}. \tag{5.12}$$

Proof Let us bound the distances $r_k = \beta_d(x_k, x_*)$ for $k \geq k(N)$. If $k \notin \mathcal{F}_N$, then

$$r_{k+1} = \beta_d(y_k, x_*) \overset{(3.11)}{\leq} r_k - \beta(x_k, y_k) \overset{(5.5)}{\leq} r_k - \tfrac{1}{2} h_k^2.$$

If $k \in \mathcal{F}_N$, then

$$r_{k+1} \overset{(2.17)}{\leq} \beta(y_k, x_*) - h_k \delta_k + \tfrac{1}{2} h_k^2 \overset{(3.11)}{\leq} r_k - h_k \delta_k + \tfrac{1}{2} h_k^2 \leq r_k - h_k \delta_N^* + \tfrac{1}{2} h_k^2.$$

Summing up these inequalities for $k = k(N), \ldots N - 1$, we obtain

$$\begin{aligned} r_N &< D - \frac{1}{2} \sum_{k \notin \mathcal{F}_N} h_k^2 + \frac{1}{2} \sum_{k \in \mathcal{F}_N} (h_k^2 - 2 h_k \delta_N^*) \\ &= D - \frac{1}{2} \sum_{k=k(N)}^{N-1} h_k^2 + \sum_{k \in \mathcal{F}_N} (h_k^2 - h_k \delta_N^*) \\ &\overset{(5.7),(5.10)}{<} \sum_{k \in \mathcal{F}_N} h_k (h_k - \delta_N^*) \overset{(5.6)_a}{\leq} (h_{k(N)} - \delta_N^*) \sum_{k \in \mathcal{F}_N} h_k. \end{aligned}$$

Thus, $\delta_N^* < h_{k(N)}$. Finally, inclusion $k \in \mathcal{F}_N$ implies

$$\tfrac{1}{2} \|y_k - x_k\|^2 \overset{(1.8)}{\leq} \beta_d(x_k, y_k) \leq \tfrac{1}{2} h_k^2 \overset{(5.6)_a}{\leq} \tfrac{1}{2} h_{k(N)}^2.$$

Since $F_* = 0$, we get

$$\begin{aligned} F(y_k) &\leq \ell_{x_k}(y_k) + \frac{L_\nu}{1+\nu} \|y_k - x_k\|^{1+\nu} + \psi(y_k) \\ &\overset{(3.12)}{\leq} \frac{L_\nu}{1+\nu} \|y_k - x_k\|^{1+\nu} \leq \frac{L_\nu}{1+\nu} h_{k(N)}^{1+\nu}, \end{aligned}$$

and this is inequality (5.12). □

As a straightforward consequence of Theorem 3, we have the following rate of convergence in function value:

$$\min_{x \in \mathcal{F}_N} f_0(y_k) \leq f_0^* + \mu(h_{k(N)}), \tag{5.13}$$

where the function $\mu(\cdot)$ is defined by (2.4).

Thus, the actual rate of convergence of the method (5.5) depends on the rate of convergence of sequence $\mathcal{T}$ and the magnitude of divergence delay. For example, for the choice (5.9), in view of inequality $a(k) \leq k-1$, we have

$$k(N) = \max\left\{k : \; k + a(k) \leq N-1\right\} \; \geq \; \max\left\{k : \; 2k-1 \leq N-1\right\}.$$

Hence, for $N = 2M$, the choice (5.9) ensures $k(N) \geq M = N/2$, and we have

$$h_{k(N)} \leq h_M \; = \; \tfrac{2D^{1/2}}{\sqrt{M+1}}. \tag{5.14}$$

It is interesting that the rate of convergence (5.12) for the constraints can be higher than the rate (5.13) for the objective function.

6 Approximating Lagrange Multipliers, I

Despite the good convergence rate (5.13), when the number of functional components m in problem (5.1) is big, the implementation of one iteration (5.3) can be very expensive. In this section, we consider a simpler switching strategy for solving convex optimization problems with potentially many functional constraints. Our method is also able to approximate the corresponding optimal Lagrange multipliers.

Consider the following *constrained* optimization problem:

$$f_0^* \; = \; \min_{x \in Q}\left\{f_0(x) : \; f_i(x) \leq 0, \; i = 1, \ldots, m\right\}, \tag{6.1}$$

where all functions $f_i(\cdot)$ are closed and convex, $i = 0, \ldots, m$, and set Q is closed, convex, and bounded. Denote by x^* one of its optimal solutions. Sometimes we use notation $\bar{f}(x) = (f_1(x), \ldots, f_m(x))^T \in \mathbb{R}^m$. Denote $\mathcal{F} = \{x \in Q : \; \bar{f}(x) \leq 0\}$.

We assume that functions $f_i(\cdot)$ are subdifferentiable on Q and it is possible to compute their subgradients with uniformly bounded norms:

$$\|f_i'(x)\|_* \leq M_i, \quad x \in Q, \quad i = 0, \ldots, m. \tag{6.2}$$

Denote $\bar{M} = (M_1, \ldots, M_m)^T \in \mathbb{R}^m$.

For the set Q, we assume existence of a prox-function $d(\cdot)$ defining the corresponding Bregman distance $\beta_d(\cdot, \cdot)$. In our methods, we need to know a constant D such that

$$\beta_d(x, y) < D, \quad \forall x, y \in Q. \tag{6.3}$$

Let us introduce the Lagrangian

$$\mathcal{L}(x, \bar{\lambda}) = f_0(x) + \langle \bar{\lambda}, \bar{f}(x) \rangle, \quad \bar{\lambda} = (\lambda^{(1)}, \ldots, \lambda^{(m)})^T \in \mathbb{R}^m_+,$$

and the dual function $\phi(\bar{\lambda}) = \min_{x \in Q} \mathcal{L}(x, \bar{\lambda})$. By Sion's theorem, we know that

$$\sup_{\bar{\lambda} \in \mathbb{R}^m_+} \phi(\bar{\lambda}) = f_0^*. \tag{6.4}$$

Our first method is based on two operations presented in Sects. 2 and 3. For defining iterations of the first type, we need functions $\varphi_{i,x}(\lambda)$ with $\lambda \geq 0$, parameterized by $x \in Q$, and defined as follows:

$$\varphi_{i,x}(\lambda) = \max_{T \in Q} \left\{ \lambda \langle f_i'(x), x - T \rangle - \beta_d(x, T) \right\}, \quad i = 0, \ldots, m. \tag{6.5}$$

An appropriate value of λ can be found from the equation

$$\varphi_{i,x_k}(\lambda) = \tfrac{1}{2} h_k^2, \tag{6.6}$$

and used for setting $x_{k+1} = T_{i,x_k}(\lambda)$, the optimal solutions of problem (6.5). In this section, we perform this iteration only for the objective function ($i = 0$). The possibility of using the steps $T_{i,x_k}(\lambda)$ for inequality constraints is analyzed in Sects. 7 and 9.

The iteration of the second type is trying to improve feasibility of the current point $x_k \in Q$ (compare with (5.3)). It needs computation of all Bregman projections

$$\hat{T}_i(x_k) \stackrel{\text{def}}{=} \arg\min_{T \in Q} \left\{ \beta_d(x_k, T) : \; f_i(x_k) + \langle f_i'(x_k), T - x_k \rangle \leq 0 \right\}, \tag{6.7}$$

for indexes $i = 1, \ldots, m$. The first-order optimality condition for problem (6.7) is as follows:

$$\langle \nabla d(T) - \nabla d(x_k) + \lambda f_i'(x_k), x - T \rangle \geq 0, \quad \forall x \in Q, \tag{6.8}$$

where $T = \hat{T}_i(x_k)$ and $\lambda = \lambda_i(x_k) \geq 0$ is the optimal Lagrange multiplier for the linear inequality constraint in (6.7). We assume that this multiplier can be also computed. Note that for $x = x_k$ we have

$$\lambda_i(x_k) \langle f_i'(x_k), x_k - T \rangle \geq \beta_d(x_k, T) + \beta_d(T, x_k) \stackrel{(1.8)}{\geq} \|x_k - T\|^2.$$

Hence, if $T \neq x_k$, then $\lambda_i(x_k) > 0$.

Consider the following optimization scheme.

Subgradient Method for Problem (6.1)

Initialization. Choose $x_0 \in Q$ and a sequence of step bounds $\mathcal{H} = \{h_k\}_{k \geq 0}$.

k**th iteration** ($k \geq 0$)**. 1.** Compute $r_k^i = \beta_d(x_k, \hat{T}_i(x_k))$, $1 \leq i \leq m$.

2. If $\exists\, i \in \{1, \ldots, m\} : r_k^i \geq \frac{1}{2} h_k^2$, then set $i_k = i$, $\lambda_k = \lambda_i(x_k)$, $x_{k+1} = \hat{T}_i(x_k)$.

3. Else, set $i_k = 0$, compute λ_k by (6.6) and set $x_{k+1} = T_{x_k}(\lambda_k)$.

(6.9)

Let us prove that method (6.9) can find an approximate solution of the primal-dual problem (6.1), (6.4). Let us choose a sequence $\mathcal{T}$ satisfying condition (5.6) and define

$$h_k = \sqrt{2D}\, \tau_k, \quad k \geq 0, \tag{6.10}$$

where D satisfies inequality (6.3). Then, for the number of steps $N \geq 1 + a(0)$, let us define function $k(N)$ by the first equation in (5.11). Now we can define the following objects:

$$\begin{aligned} \mathcal{A}_i(N) &= \{k : \ k(N) \leq k \leq N-1, \ i_k = i\}, \\ \sigma_i(N) &= \sum_{k \in \mathcal{A}_i(N)} \lambda_k, \quad i = 0, \ldots, m, \\ \lambda_i^*(N) &= \sigma_i(N)/\sigma_0(N), \quad i = 1, \ldots, m. \end{aligned} \tag{6.11}$$

Clearly, the vectors of dual multipliers $\bar{\lambda}_*(N) = (\lambda_1^*(N), \ldots, \lambda_m^*(N))$ are defined only if $\mathcal{A}_0(N) \neq \emptyset$. As usual, the sum over an empty set of iterations is assumed to be zero.

Theorem 4 *Let the sequence of step bounds $\mathcal{H}$ in method (6.9) be defined by (6.10). Then for any $N \geq 1 + a(0)$, we have $\mathcal{A}_0(N) \neq \emptyset$. Moreover, if all functions $f_i(\cdot)$, $0 = 1, \ldots, m$, satisfy (6.2), then*

$$\max_{1 \leq i \leq m} \tfrac{1}{M_i} f_i(x_k) \leq h_{k(N)}, \quad k \in \mathcal{A}_0(N), \tag{6.12}$$

$$\tfrac{1}{\sigma_0(N)} \sum_{k \in \mathcal{A}_0(N)} \lambda_k f_0(x_k) \leq \phi(\bar{\lambda}^*(N)) + M_0 h_{k(N)}. \tag{6.13}$$

Proof Indeed, we have

$$B_N \stackrel{\text{def}}{=} \max_{x \in Q} \Big[\sum_{k \in \mathcal{A}_0(N)} \lambda_k f_0(x_k) - \sigma_0(N) f_0(x) - \sum_{i=1}^{m} \sigma_i(N) f_i(x) \Big]$$

$$= \max_{x \in Q} \Big[\sum_{k \in \mathcal{A}_0(N)} \lambda_k [f_0(x_k) - f_0(x)] - \sum_{i=1}^{m} \sigma_i(N) f_i(x) \Big].$$

Let us fix some $x \in Q$ and denote $r_k(x) = \beta_d(x_k, x)$. Then, for $k \in \mathcal{A}_0(N)$, we have

$$\lambda_k [f_0(x_k) - f_0(x)] \le \lambda_k \langle f_0'(x_k), x_k - x \rangle. \tag{6.14}$$

At the same time, from the first-order optimality condition for problem (6.5), we have

$$0 \le \langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle + \lambda_k \langle f_0'(x_k), x - x_{k+1} \rangle$$

$$\stackrel{(6.14)}{\le} \langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle - \lambda_k [f_0(x_k) - f_0(x)] + \lambda_k \langle f_0'(x_k), x_k - x_{k+1} \rangle$$

$$= \langle \nabla d(x_{k+1}) - \nabla d(x_k), x - x_{k+1} \rangle - \lambda_k [f_0(x_k) - f_0(x)] + \varphi_{0,x_k}(\lambda_k) + \beta_d(x_k, x_{k+1}).$$

Hence,

$$r_{k+1}(x) - r_k(x) \;=\; \langle \nabla d(x_k) - \nabla d(x_{k+1}), x - x_{k+1} \rangle - \beta_d(x_k, x_{k+1})$$

$$\le -\lambda_k [f_0(x_k) - f_0(x)] + \varphi_{0,x_k}(\lambda_k) \stackrel{(6.6)}{=} -\lambda_k [f_0(x_k) - f_0(x)] + \tfrac{1}{2} h_k^2.$$

On the other hand,

$$\sum_{i=1}^{m} \sigma_i(N) f_i(x) \stackrel{(6.11)}{=} \sum_{i=1}^{m} \sum_{k \in \mathcal{A}_i(N)} \lambda_k f_i(x) \;=\; \sum_{k \notin \mathcal{A}_0(N)} \lambda_{i_k}(x_k) f_{i_k}(x)$$

$$\ge \sum_{k \notin \mathcal{A}_0(N)} \lambda_{i_k}(x_k) [f_{i_k}(x_k) + \langle f_{i_k}'(x_k), x - x_k \rangle]$$

$$\stackrel{(6.7)}{=} \sum_{k \notin \mathcal{A}_0(N)} \lambda_{i_k}(x_k) [\langle f_{i_k}'(x_k), x - x_{k+1} \rangle]$$

$$\stackrel{(6.8)}{\ge} \sum_{k \notin \mathcal{A}_0(N)} \langle \nabla d(x_{k+1}) - \nabla d(x_k), x_{k+1} - x \rangle$$

$$= \sum_{k \notin \mathcal{A}_0(N)} \Big[r_{k+1}(x) - r_k(x) + \beta_d(x_k, x_{k+1}) \Big]$$

$$\stackrel{(6.9)_2}{\geq} \sum_{k \notin \mathcal{A}_0(N)} \Bigl[r_{k+1}(x) - r_k(x) + \frac{1}{2} h_k^2 \Bigr].$$

Therefore,

$$\begin{aligned} B_N &\leq \max_{x \in Q} \Bigl[\sum_{k \in \mathcal{A}_0(N)} [r_k(x) - r_{k+1}(x) + \tfrac{1}{2} h_k^2] - \sum_{k \notin \mathcal{A}_0(N)} [r_{k+1}(x) - r_k(x) + \tfrac{1}{2} h_k^2] \Bigr] \\ &= \max_{x \in Q} \Bigl[r_{k(N)}(x) - r_N(x) + \tfrac{1}{2} \sum_{k \in \mathcal{A}_0(N)} h_k^2 - \tfrac{1}{2} \sum_{k \notin \mathcal{A}_0(N)} h_k^2 \Bigr] \\ &< D - \tfrac{1}{2} \sum_{k=k(N)}^{N-1} h_k^2 + \sum_{k \in \mathcal{A}_0(N)} h_k^2 \stackrel{(6.10)}{\leq} \sum_{k \in \mathcal{A}_0(N)} h_k^2 \stackrel{(5.6)_a}{\leq} h_{k(N)} \sum_{k \in \mathcal{A}_0(N)} h_k. \end{aligned}$$

Let us assume now that $\mathcal{A}_0(N) = \emptyset$. Then $B_N < 0$, and this is impossible since the point x_* is feasible. Thus, we have proved that $\mathcal{A}_0(N) \neq \emptyset$.

Further, for any $k \in \mathcal{A}_0(N)$ we have

$$\lambda_k \langle f_0'(x_k), x_k - x_{k+1} \rangle \stackrel{(6.6)}{=} \beta_d(x_k, x_{k+1}) + \tfrac{1}{2} h_k^2.$$

Hence, $r_k^0 = \|x_{k+1} - x_k\| > 0$ and we conclude that

$$\lambda_k M_0 \stackrel{(6.2)}{\geq} \lambda_k \|f_0'(x_k)\|_* \geq \tfrac{\lambda_k}{r_k^0} \langle f_0'(x_k), x_k - x_{k+1} \rangle \stackrel{(1.8)}{\geq} \tfrac{1}{2r_k^0} \Bigl[(r_k^0)^2 + h_k^2 \Bigr] \geq h_k.$$

Thus, we have proved that $B_N \leq M_0 \sigma_0(N) h_{k(N)}$. Dividing this inequality by $\sigma_0(N) > 0$, we get the bound (6.13).

Finally, note that for any $k \in \mathcal{A}_0(N)$ we have $\|\hat{T}_i(x_k) - x_k\| \leq h_k$ for all i, $1 \leq i \leq m$. Note that either $f_i(x_k) \leq 0$, or $f_i(x_k) > 0$ and $f_i(x_k) + \langle f_i'(x_k), \hat{T}_i(x_k) - x_k) \rangle = 0$. In the latter case, we have

$$f_i(x_k) = \langle f_i'(x_k), x_k - \hat{T}_i(x_k) \rangle \stackrel{(6.2)}{\leq} M_i h_k \leq M_i h_{k(N)}.$$

Thus, inequality (6.12) is proved. □

Note that for the choice of scaling sequence (5.9), method (5.5) is globally convergent. In this case, the computation of Lagrange multipliers by (6.11) for all values of N, requires storage of all coefficients $\{\lambda_k\}_{k \geq 0}$. This inconvenience can be avoided if we decide to accumulate the sums for Lagrange multipliers only starting from the moments $k(N_q)$ with $N_q = 2^q$, $q \geq 1$. Then the method (5.5) will be allowed to stop only at the moments $2N_q$.

7 Approximating Lagrange Multipliers, II

Method (6.9) has one hidden drawback. If $i_k \geq 1$, then

$$0 \overset{(6.7)}{=} f_{i_k}(x_k) + \langle f'_{i_k}(x_k), x_{k+1} - x_k \rangle \leq f_{i_k}(x_{k+1}).$$

Thus, this scheme most probably generates infeasible approximations of the optimal point, which violate some of the functional constraints. In order to avoid this tendency, we propose a scheme which uses for both types of iterations (improving either feasibility or optimality) the same step-size rule (6.6).

Fixed-Step Switching Subgradient Method for Problem (6.1)

Initialization. Choose $x_0 \in Q$ and a sequence of step bounds $\mathcal{H} = \{h_k\}_{k \geq 0}$.

k**th iteration** ($k \geq 0$)**. 1.** Compute $T_{i,x_k}(\lambda_{i,k})$ by (6.6), $1 \leq i \leq m$.

2. If $\exists\, i \in \{1, \ldots, m\} : \lambda_{i,k} f_i(x_k) \geq h_k^2$, then set $i_k = i$ and $x_{k+1} = T_{i,x_k}(\lambda_{i,k})$.

3. Else, set $i_k = 0$, compute $T_{0,x_k}(\lambda_{0,k})$ by (6.6), and set $x_{k+1} = T_{0,k}(\lambda_{0,k})$.

4. Set $\lambda_k = \lambda_{i_k,k}$.

(7.1)

Thus, for both types of iterations, we use the same step-size strategy (6.6). Note that for any $i = 0, \ldots, m$ and $T_i = T_{i,x_k}(\lambda_{i,k})$, we have

$$\lambda_{i,k} \langle f'_i(x_k), x_k - T_i \rangle \overset{(6.5)}{=} \varphi_{i,x_k}(\lambda_{i,k}) + \beta_d(x_k, T_i) \overset{(6.6)}{=} \tfrac{1}{2} h_k^2 + \beta_d(x_k, T_i)$$

$$\overset{(1.8)}{\geq} \tfrac{1}{2} h_k^2 + \tfrac{1}{2} \|x_k - T_i\|^2 \geq h_k \|x_k - T_i\|.$$

Hence, since $T_i \neq x_k$, we have

$$\lambda_{i,k} \|f'_i(x_k)\|_* \geq h_k, \quad k \geq 0. \tag{7.2}$$

In method (7.1), we choose $\mathcal{H}$ in accordance to (6.10). Then, for $N \geq 1 + a(0)$, we define function $k(N)$ by the first equation in (5.11) and introduce by (6.11) the approximations of the optimal Lagrange multipliers.

Theorem 5 *Let the sequence of points $\{x_k\}_{k \geq 0}$ be generated by method (7.1). Then for any $N \geq 1 + a(0)$ the set $\mathcal{A}_0(N)$ is not empty. Moreover, if all functions $f_i(\cdot)$,*

$i = 1, \ldots, m$, satisfy (6.2), then for all $k \in \mathcal{A}_0(N)$ we have

$$f_i(x_k) \le \|f_i'(x_k)\|_* h_k \overset{(6.2)}{\le} M_i h_{k(N)}, \quad i = 1, \ldots, m, \tag{7.3}$$

$$\tfrac{1}{\sigma_0(N)} \sum_{k \in \mathcal{A}_0(N)} \lambda_k f_0(x_k) \le \phi(\bar{\lambda}^*(N)) + M_0 h_{k(N)}. \tag{7.4}$$

Proof As in the proof of Theorem 4, for $x \in Q$, we denote $r_k(x) = \beta_d(x_k, x)$, and

$$B_N \overset{\text{def}}{=} \max_{x \in Q} \Big[\sum_{k \in \mathcal{A}_0(N)} \lambda_k f_0(x_k) - \sigma_0(N) f_0(x) - \sum_{i=1}^m \sigma_i(N) f_i(x) \Big]$$

$$= \max_{x \in Q} \Big[\sum_{k \in \mathcal{A}_0(N)} \lambda_k [f_0(x_k) - f_0(x)] - \sum_{i=1}^m \sigma_i(N) f_i(x) \Big].$$

Then, for $k \in \mathcal{A}_0(N)$, we have proved that

$$r_{k+1}(x) - r_k(x) \le -\lambda_k [f_0(x_k) - f_0(x)] + \tfrac{1}{2} h_k^2.$$

On the other hand,

$$\sum_{i=1}^m \sigma_i(N) f_i(x) \overset{(6.11)}{=} \sum_{i=1}^m \sum_{k \in \mathcal{A}_i(N)} \lambda_k f_i(x) = \sum_{k \notin \mathcal{A}_0(N)} \lambda_{i_k,k} f_{i_k}(x)$$

$$\ge \sum_{k \notin \mathcal{A}_0(N)} \lambda_{i_k,k} [f_{i_k}(x_k) + \langle f_{i_k}'(x_k), x - x_k \rangle]$$

$$\overset{(7.1)_2}{\ge} \sum_{k \notin \mathcal{A}_0(N)} \Big[h_k^2 + \lambda_{i_k,k} \langle f_{i_k}'(x_k), x - x_k \rangle \Big].$$

Note that the first-order optimality condition for problem (2.7) is as follows:

$$\langle \nabla d(T_i) - \nabla d(x_k) + \lambda f_i'(x_k), x - T_i \rangle \ge 0, \quad \forall x \in Q,$$

where T_i is its optimal solution, Therefore, for all $x \in Q$ an $k \notin \mathcal{A}_0(N)$, we have

$$\lambda_{i_k,k} \langle f_{i_k}'(x_k), x - x_k \rangle = \lambda_{i_k,k} \Big[\langle f_{i_k}'(x_k), x - x_{k+1} \rangle + \langle f_{i_k}'(x_k), x_{k+1} - x_k \rangle \Big]$$

$$\ge \lambda_{i_k,k} \langle f_{i_k}'(x_k), x_{k+1} - x_k \rangle + \langle \nabla d(x_{k+1}) - \nabla d(x_k), x_{k+1} - x \rangle$$

$$= \lambda_{i_k,k} \langle f_{i_k}'(x_k), x_{k+1} - x_k \rangle + r_{k+1}(x) - r_k(x) + \beta_d(x_k, x_{k+1})$$

$$= r_{k+1}(x) - r_k(x) - \varphi_{i_k,x_k}(\lambda_{i,k}) \overset{(6.6)}{=} r_{k+1}(x) - r_k(x) - \tfrac{1}{2} h_k^2.$$

MThus, as in the proof of Theorem 4, we conclude that

$$B_N \le \max_{x \in Q} \Big[\sum_{k \in \mathcal{A}_0(N)} [r_k(x) - r_{k+1}(x) + \tfrac{1}{2} h_k^2] - \sum_{k \notin \mathcal{A}_0(N)} [r_{k+1}(x) - r_k(x) + \tfrac{1}{2} h_k^2] \Big]$$

$$\le h_{k(N)} \sum_{k \in \mathcal{A}_0(N)} h_k \overset{(7.2)}{\le} M_0 h_{k(N)} \sigma_0(N).$$

Since $\mathcal{A}_0(N) \neq \emptyset$ (see the proof of Theorem 4), dividing this inequality by $\sigma_0(N) > 0$, we get inequality (7.4).

Finally, note that for all $k \in \mathcal{A}_0(N)$ and $i = 1, \dots, m$, we have

$$\lambda_{i,k} f_i(x_k) \le h_k^2 \overset{(7.2)}{\le} \lambda_{i,k} \| f_i'(x_k) \|_* h_k \overset{(6.2)}{\le} \lambda_{i,k} M_i h_k \le \lambda_{i,k} M_i h_{k(N)}.$$

Thus, inequality (7.3) is proved. □

8 Accuracy guarantees for the dual problem

In Sects. 6 and 7, we developed two convergent methods (6.9) and (7.1), which are able to approach the optimal solution of the primal problem (6.1), generating in parallel an approximate solution of the dual problem (6.4). Indeed, for $N \ge 1 + a(0)$, denote

$$f_0^*(N) = \min_{k \in \mathcal{A}_0(N)} f_0(x_k), \quad x_N^* = \arg\min_x \Big\{ f_0(x) : \; x = x_k, \; k \in \mathcal{A}_0(N) \Big\}.$$

Then, in view of Theorems 4 and 5, we have

$$\begin{aligned} f_0^*(N) - \phi(\bar{\lambda}^*(N)) &\le M_0 h_{k(N)}, \\ \max_{1 \le i \le m} \tfrac{1}{M_i} f_i(x_k) &\le h_{k(N)}, \quad k \in \mathcal{A}_0(N). \end{aligned} \tag{8.1}$$

Since $\phi(\bar{\lambda}^*(N)) \le f_0^*$, this inequality justfies that the point x_N^* is a good approximate solution to primal problem (6.1).

However, note that in our reasonings, we did not assume yet the *existence* of an optimal solution to the dual problem (6.4). It appears that under our assumptions, this may not happen. In this case, since $f_0^*(N)$ can be *significantly smaller* than f_0^*, inequality (8.1) cannot justify that vector $\bar{\lambda}^*(N)$ delivers a good value of the dual objective function.

Let us look at the following example.

Example 4 Consider the following problem

$$\min_{x \in \mathbb{R}^2} \Big\{ x^{(2)} : \; 1 - x^{(1)} \le 0, \; \|x\|_2 \le 1 \Big\}.$$

In this case, $\mathcal{L}(x, \lambda) = x^{(2)} + \lambda(1 - x^{(1)})$. Hence,

$$\phi(\lambda) = \min_{\|x\|_2 \le 1} \mathcal{L}(x, \lambda) \;=\; \lambda - [1 + \lambda^2]^{1/2}.$$

Thus, there is no duality gap: $\phi_* \stackrel{\text{def}}{=} \sup_{\lambda \ge 0} \phi(\lambda) = 0$. However, the optimal dual solution λ^* does not exist.

Let us look now at the perturbed feasible set

$$\mathcal{F}()\epsilon) = \{x \in \mathbb{R}^2 : \; \|x\|_2 \le 1, \; 1 - \epsilon \le x^{(1)}\},$$

where $\epsilon > 0$ is sufficiently small. Note that it contains a point with the second coordinate equal to $\phi_* - \sqrt{\epsilon(2 - \epsilon)}$. This means that the condition (8.1) can guarantee only that

$$\phi(\lambda^*(N) \ge \phi_* - \epsilon - \sqrt{\epsilon(2 - \epsilon)}.$$

Hence, for dual problems with nonexisting optimal solutions, we can expect a significant drop in the quality of approximation in terms of the function value. □

Thus, in our complexity bounds, we need to take into account the size of optimal dual solution $\bar{\lambda}^* \in \mathbb{R}^m_+$. Let the sequence $\{x_k\}_{k \ge 0}$ be generated by one of the methods (6.9) or (7.1). Then, for any $k \ge 0$, we have

$$\begin{aligned} f_0(x_k) &\ge \min_{x \in Q} \left\{ f_0(x) : \; \bar{f}(x) \le \bar{f}(x_k) \right\} = \min_{x \in Q} \max_{\bar{\lambda} \in \mathbb{R}^m_+} \left\{ f_0(x) + \langle \bar{\lambda}, \bar{f}(x) - \bar{f}(x_k) \rangle \right\} \\ &= \max_{\bar{\lambda} \in \mathbb{R}^m_+} \min_{x \in Q} \left\{ f_0(x) + \langle \bar{\lambda}, \bar{f}(x) - \bar{f}(x_k) \rangle \right\} \;=\; \max_{\bar{\lambda} \in \mathbb{R}^m_+} \left\{ \phi(\bar{\lambda}) - \langle \bar{\lambda}, \bar{f}(x_k) \rangle \right\} \\ &\ge f_0^* - \langle \bar{\lambda}^*, \bar{f}(x_k) \rangle \;\ge\; f_0^* - \langle \bar{\lambda}^*, \bar{M} \rangle \max_{1 \le i \le m} \tfrac{1}{M_i} f_i(x_k). \end{aligned} \tag{8.2}$$

Thus, we have proved the following theorem.

Theorem 6 *Under conditions of Theorem 4 or 5, for all $N \ge 1 + a(0)$ and all $k \in \mathcal{A}_0(N)$, we have*

$$f_0^* - \phi(\bar{\lambda}^*(N)) \le (M_0 + \langle \bar{\lambda}^*, \bar{M} \rangle)\, h_{k(N)}. \tag{8.3}$$

Proof Indeed, in view of inequalities (6.12) and (7.3), we have $\max_{1 \le i \le m} \frac{1}{M_i} f_i(x_k) \le h_{k(N)}$. Thus, we get the bound (8.3) from (8.1) and (8.2). □

Recall that the size of optimal dual multipliers can be bounded by the standard *Slater condition*. Namely, let us assume existence of a point $\hat{x} \in Q$ such that

$$f_i(\hat{x}) < 0, \quad i = 1, \ldots, m. \tag{8.4}$$

Then, in accordance to Lemma 3.1.21 in [5], we have

$$\langle \bar{\lambda}^*, -\bar{f}(\hat{x}) \rangle \leq f_0(\hat{x}) - f_0^*. \tag{8.5}$$

Therefore,

$$\langle \bar{\lambda}^*, \bar{M} \rangle = \sum_{i=1}^{m} \lambda_i^* (-f_i(\hat{x})) \cdot \tfrac{M_i}{-f_i(\hat{x})} \leq \langle \bar{\lambda}^*, -\bar{f}(\hat{x}) \rangle \max_{1 \leq i \leq m} \tfrac{M_i}{-f_i(\hat{x})}$$

$$\overset{(8.5)}{\leq} (f_0(\hat{x}) - f_0^*) \max_{1 \leq i \leq m} \tfrac{M_i}{-f_i(\hat{x})}.$$

Thus, the Slater condition provides us with the following bound:

$$f_0^* - \phi(\bar{\lambda}^*(N)) \leq \left(M_0 + (f_0(\hat{x}) - f_0^*) \max_{1 \leq i \leq m} \tfrac{M_i}{-f_i(\hat{x})} \right) h_{k(N)}, \tag{8.6}$$

which is valid for all $N \geq 1 + a(0)$. Note that we are able to compute vector $\bar{\lambda}^*(N)$ without computing values of the dual function $\phi(\cdot)$, which can be very complex. In fact, computational complexity of a single value $\phi(\lambda)$ can be of the same order as the complexity of solving the initial problem (6.1), or even more.

9 Subgradient method for unbounded feasible set

In the previous sections, we looked at optimization methods applicable to the bounded sets (see condition (6.3)). If this is not true, the second-order divergence condition (5.6) cannot help, and we need to find another way of justifying efficiency of the subgradient schemes. This is the goal of the current section.

We are still working with the problem (6.1), satisfying condition (6.2). However, the set Q is not bounded anymore. Hence, we cannot count on Sion's theorem (6.4).

In our method, we have a sequence of scaling coefficients $\Gamma = \{\gamma_k\}_{k \geq 0}$ satisfying condition

$$\gamma_{k+1} > \gamma_k \geq 0, \quad k \geq 0, \tag{9.1}$$

a tolerance parameter $\epsilon > 0$, and a rough estimate for the distance to the optimum $D_0 > 0$.

For functions $\varphi_{i,y}(\cdot)$ defined by (6.5) with $y \in Q$, denote by $a = a_{i,k}(y)$ the unique solution of the equation

$$\varphi_{i,y}\left(\tfrac{a}{\gamma_{k+1}} \right) = \left(1 - \tfrac{\gamma_k}{\gamma_{k+1}} \right) D_0, \quad i = 0, \ldots, m. \tag{9.2}$$

Let us look at the following optimization scheme.

Switching Subgradient Method for Unbounded Sets

Initialization. Choose $x_0 \in Q$, a sequence Γ (see (9.1)), and some $D_0 > 0$.

kth iteration ($k \geq 0$)**. 1.** Define $\tau_k = \frac{\gamma_k}{\gamma_{k+1}}$ and $y_k = (1 - \tau_k)x_0 + \tau_k x_k$.

2. If $\exists\, i \in \{1, \ldots, m\} : \; f_i(y_k) \geq \epsilon$, then set $i_k = i$. Otherwise, set $i_k = 0$.

3. Compute $a_k = a_{i_k,k}(y_k)$ by (9.2), and set $x_{k+1} = T_{i_k,y_k}\left(\frac{a_k}{\gamma_{k+1}}\right)$.

(9.3)

It seems that now, the selection of violated constraint by Step 2 looks more natural than in the methods (6.9) and (7.1).

For $x \in Q$, denote $\Delta_k(x) = \gamma_k\Big(\beta_d(x_k, x) - \beta_d(x_0, x)\Big)$. This is a linear function of x. In what follows, we assume the Bregman distance is convex with respect to its first argument:

$$\beta_d(\alpha u_1 + (1-\alpha)u_2, x) \leq \alpha\beta_d(u_1, x) + (1-\alpha)\beta_d(u_2, x), \qquad \forall u_1, u_2, x \in Q,\ \alpha \in [0, 1]. \tag{9.4}$$

This property is not very common. However, it is valid for the following two important examples.

Example 5

1. Let $d(x) = \frac{1}{2}\|x\|_B^2$ (see (1.5)). Then $\beta_d(x, y) = \frac{1}{2}\|x - y\|_B^2$ and (9.4) holds.

2. Let $Q = \{x \in \mathbb{R}^n_+ : \sum\limits_{i=1}^n x^{(i)} = 1\}$. Define $d(x) = \sum\limits_{i=1}^n x^{(i)} \ln x^{(i)}$. Then, for $x, y \in Q$, we have $\beta_d(x, y) = \sum\limits_{i=1}^n y^{(i)} \ln \frac{y^{(i)}}{x^{(i)}}$, and (9.4) holds also. □

Now we can prove the following statement.

Lemma 6 *Let the Bergman distance $\beta_d(\cdot, \cdot)$ satisfy (9.4). Then, for the sequence $\{x_k\}_{k\geq 0}$ generated by method (9.3), and any $x \in Q$, we have*

$$\Delta_{k+1}(x) \leq \Delta_k(x) + a_k\langle f'_{i_k}(y_k), x - y_k\rangle + (\gamma_{k+1} - \gamma_k)D_0, \quad k \geq 0. \tag{9.5}$$

Proof Note that

$$\begin{aligned}\Delta_{k+1}(x) - \Delta_k(x) &= \gamma_{k+1}\Big(\beta_d(x_{k+1}, x) - \beta_d(x_0, x)\Big) - \gamma_k\Big(\beta_d(x_k, x) - \beta_d(x_0, x)\Big) \\ &\overset{(9.4)}{\leq} \gamma_{k+1}\Big(\beta_d(x_{k+1}, x) - \beta_d(y_k, x)\Big) \\ &\overset{(1.6)}{=} \gamma_{k+1}\Big(\langle \nabla d(y_k) - \nabla d(x_{k+1}), x - x_{k+1}\rangle - \beta_d(y_k, x_{k+1})\Big).\end{aligned}$$

Since the first-order optimality conditions for point x_{k+1} tells us that

$$\langle a_k f'_{i_k}(y_k) + \gamma_{k+1}(\nabla d(x_{k+1}) - \nabla d(y_k)), x - x_{k+1}\rangle \geq 0, \quad x \in Q,$$

we have

$$\begin{aligned}\Delta_{k+1}(x) - \Delta_k(x) &\leq a_k\langle f'_{i_k}(y_k), x - x_{k+1}\rangle - \gamma_{k+1}\beta_d(y_k, x_{k+1}\rangle \\ &\overset{(6.5)}{=} a_k\langle f'_{i_k}(y_k), x - y_k\rangle + \gamma_{k+1}\varphi_{i,y_k}\left(\tfrac{a_k}{\gamma_{k+1}}\right) \\ &\overset{(9.2)}{=} a_k\langle f'_{i_k}(y_k), x - y_k\rangle + (\gamma_{k+1} - \gamma_k)D_0,\end{aligned}$$

and this is inequality (9.5). □

Since $\Delta_0(x) = 0$, we can sum up the inequalities (9.5) for $k = 0, \ldots, N-1$ with $N \geq 1$, and get the following consequence:

$$\gamma_N\beta_d(x_N, x) + \sum_{k=0}^{N-1} a_k\langle f'_{i_k}(y_k), y_k - x\rangle \overset{(9.5)}{\leq} \gamma_N(\beta_d(x_0, x) + D_0), \tag{9.6}$$

which is valid for all $x \in Q$.

In order to approximate the optimal Lagrange multipliers of problem (6.4), we need to introduce the following objects:

$$\begin{aligned}&\mathcal{S}_i(N) = \{k : \ 0 \leq k \leq N-1, \ i_k = i\}, \quad N \geq 1, \\ &\hat{\sigma}_i(N) = \sum_{k \in \mathcal{S}_i(N)} a_k, \quad i = 0, \ldots, m, \\ &\hat{\lambda}_i(N) = \hat{\sigma}_i(N)/\hat{\sigma}_0(N), \quad i = 1, \ldots, m.\end{aligned} \tag{9.7}$$

Denote $\hat{\lambda}(N) = (\hat{\lambda}_1(N), \ldots, \hat{\lambda}_m(N))^T \in \mathbb{R}^m_+$ and $f_0^*(N) = \min\limits_{k \in \mathcal{S}_0(N)} f_0(x_k)$. Note that for all $k \in \mathcal{S}_0(N)$ we have

$$f_i(x_k) \leq \epsilon, \quad k = 1, \ldots, m. \tag{9.8}$$

For our convergence result, we need to assume existence of the optimal solution x^* to problem (6.1). Let us introduce some bound

$$D \geq \beta_d(x_0, x^*),$$

which is not used in the method (9.3). Denote $Q_D = \{x \in Q : \beta_d(x_0, x) \leq D\}$. Clearly, if we replace in problem (6.1) the set Q by Q_D, then its optimal solution will not be changed. However, now we can correctly define a *restricted* dual function $\phi_D(\lambda) = \min\limits_{x \in Q_D} \mathcal{L}(x, \lambda)$. Let us prove our main result.

Theorem 7 *Let functional components of problem* (6.1) *satisfy the following condition:*

$$\|f_i'(x)\|_* \leq M, \quad i = 0, \ldots, m. \tag{9.9}$$

Then, as far as

$$\Sigma_N \stackrel{\text{def}}{=} \frac{1}{\gamma_N} \sum_{k=0}^{N-1} \Big[\gamma_{k+1}(\gamma_{k+1} - \gamma_k)\Big]^{1/2} > \frac{\rho_0 + D_0}{\sqrt{2D_0}} \cdot \frac{M}{\epsilon}, \tag{9.10}$$

where $\rho_0 = \min\limits_{y \in \mathcal{F}} \beta_d(x_0, y)$, *we have* $\mathcal{S}_0(N) \neq \emptyset$ *and* $\hat{\sigma}_0(N) > 0$. *Moreover, if*

$$\Sigma_N > \frac{D + D_0}{\sqrt{2D_0}} \cdot \frac{M}{\epsilon}, \tag{9.11}$$

then

$$f_0^*(N) - f_0^* \;\leq\; f_0^*(N) - \phi_D(\hat{\lambda}(N)) \leq \epsilon. \tag{9.12}$$

Proof Indeed, for any $x \in Q$, we have

$$\begin{aligned}
\sum_{k=0}^{N-1} a_k \langle f_{i_k}'(y_k), y_k - x \rangle &= \sum_{k \in \mathcal{S}_0(N)} a_k \langle f_0'(y_k), y_k - x \rangle + \sum_{i=1}^{m} \sum_{k \in \mathcal{S}_i(N)} a_k \langle f_i'(y_k), y_k - x \rangle \\
&\geq \sum_{k \in \mathcal{S}_0(N)} a_k [f_0(y_k) - f_0(x)] + \sum_{i=1}^{m} \sum_{k \in \mathcal{S}_i(N)} a_k [f_i(y_k) - f_i(x)] \\
&\geq \hat{\sigma}_0(N)[f_0^*(N) - f_0(x)] + \sum_{i=1}^{m} \hat{\sigma}_i(N)[\epsilon - f_i(x)].
\end{aligned}$$

Hence, in view of inequality (9.6), we get the following bound

$$\hat{\sigma}_0(N)[f_0^*(N) - f_0(x)] + \sum_{i=1}^{m} \hat{\sigma}_i(N)[\epsilon - f_i(x)] \leq \gamma_N(\beta_d(x_0, x) + D_0).$$

Note that $\hat{\sigma}_0(N) + \sum\limits_{i=1}^{m} \hat{\sigma}_i(N) = \sum\limits_{k=0}^{N-1} a_k$. Therefore, this inequality can be rewritten as follows:

$$\gamma_N(\beta_d(x_0, x) + D_0) \geq \hat{\sigma}_0(N)[f_0^*(N) - f_0(x) - \epsilon] + \epsilon \sum_{k=0}^{N-1} a_k - \sum_{i=1}^{m} \hat{\sigma}_i(N) f_i(x). \tag{9.13}$$

In view of condition (9.2), for $T_i = T_{i,y_k}\left(\frac{a_{i,k}(y_k)}{\gamma_{k+1}}\right)$, we have

$$\begin{aligned} a_{i,k}(y_k)\langle f_i'(y_k), y_k - T_i\rangle &= \gamma_{k+1}\beta_d(y_k, T_i) + (\gamma_{k+1} - \gamma_k)D_0 \\ &\overset{(1.8)}{\geq} \frac{1}{2}\gamma_{k+1}\|T_i - y_k\|^2 + (\gamma_{k+1} - \gamma_k)D_0 \\ &\geq \left[2\gamma_{k+1}(\gamma_{k+1} - \gamma_k)D_0\right]^{1/2}\|T_i - y_k\|. \end{aligned}$$

Thus, $Ma_k(y_k) \overset{(9.9)}{\geq} \left[2\gamma_{k+1}(\gamma_{k+1} - \gamma_k)D_0\right]^{1/2}$, and we conclude that

$$\sum_{k=0}^{N-1} a_k \geq \tfrac{1}{M}\sqrt{2D_0}\gamma_N \Sigma_N. \tag{9.14}$$

Let us assume that $\mathcal{S}_0(N) = \emptyset$. Then $\hat{\sigma}_0(N) = 0$ and for $x = \arg\min\limits_{y \in \mathcal{F}} \beta_d(x_0, y)$ inequality (9.13) leads to the following relation:

$$\gamma_N(\rho_0 + D_0) \overset{(9.14)}{\geq} \tfrac{\epsilon}{M}\sqrt{2D_0}\gamma_N \Sigma_N.$$

However, this cannot happen in view of condition (9.10). Hence, inequality (9.10) implies $\mathcal{S}_0(N) \neq \emptyset$ and $\hat{\sigma}_0(N) > 0$. In this case, inequality (9.13) can be rewritten as follows:

$$\beta_d(x_0, x) + D_0 \overset{(9.14)}{\geq} \tfrac{1}{\gamma_N}\hat{\sigma}_0(N)[f_0^*(N) - \mathcal{L}(x, \hat{\lambda}(N)) - \epsilon] + \tfrac{\epsilon}{M}\sqrt{2D_0}\Sigma_N.$$

Maximizing the right-hand side of this inequality in $x \in Q_d$, we get

$$D + D_0 \geq \tfrac{1}{\gamma_N}\hat{\sigma}_0(N)[f_0^*(N) - \phi_D(\hat{\lambda}(N)) - \epsilon] + \tfrac{\epsilon}{M}\sqrt{2D_0}\Sigma_N.$$

Hence, inequality (9.11) implies (9.12). □

Thus, for the fast rate of convergence of method (9.3), we need to ensure a fast growth of the valies Σ_N. Let us choose

$$\gamma_k = \sqrt{k}, \quad k \geq 0. \tag{9.15}$$

Then $\gamma_{k+1}(\gamma_{k+1} - \gamma_k) = \frac{\sqrt{k+1}}{\sqrt{k+1}+\sqrt{k}} \geq \frac{1}{2}$. Hence,

$$\Sigma_N \geq \frac{1}{\sqrt{N}} \cdot \frac{N}{\sqrt{2}} = \sqrt{\frac{N}{2}}. \tag{9.16}$$

In this case, inequality (9.12) can be ensured in $O(\epsilon^{-2})$ iterations.

Note that method (9.3) is quite different from the existing optimization schemes. Let us write down how it looks like in the case $Q = \mathbb{E}$, $d(x) = \frac{1}{2}\|x\|_B^2$, and the parameter choice (9.15). In this case, the Eq. (9.2) can be written as follows:

$$\frac{a^2}{2\gamma_{k+1}^2}\|f_i'(y_k)\|^2 = \frac{\gamma_{k+1}-\gamma_k}{\gamma_{k+1}} D_0.$$

Thus, $a_k = \sqrt{2\gamma_{k+1}(\gamma_{k+1} - \gamma_k) D_0} \cdot \frac{1}{\|f_{i_k}'(y_k)\|_B^*} \approx \frac{\sqrt{D_0}}{\|f_{i_k}'(y_k)\|_B^*}$, and the method looks as follows:

Switching Subgradient Method With $Q = \mathbb{E}$

0. Choose $x_0 \in \mathbb{E}$ and $D_0 > 0$. **For $k \geq 0$, iterate:**

1. Define $\tau_k = \sqrt{\frac{k}{k+1}}$ and $y_k = (1 - \tau_k)x_0 + \tau_k x_k$.

2. If $\exists i : 1 \leq i \leq m$, $f_i(y_k) \geq \epsilon$, then set $i_k = i$. Otherwise, set $i_k = 0$.

3. Set $x_{k+1} = x_k - \sqrt{\frac{D_0}{k+1}} \cdot B^{-1} f_{i_k}'(y_k)/\|f_{i_k}'(y_k)\|_B^*$.

(9.17)

As compared with other known switching primal-dual schemes (e.g. Section 3.2.5 in [5]), in method (9.17), we apply variable step sizes, which allow faster movements in the beginning of the process.

10 Conclusions

In this paper, we presented several new *primal* subgradient methods with a better control of step sizes. Our main observation is that for constrained minimization problems, in a proper definition of the actual size of subgradients of convex function, we must take into account the position of the test point with respect to the boundary of the feasible set. In the simplest variant, this can be done by ensuring the size of the proximal step to be equal to a prescribed control parameter.

In this way, we develop new methods for solving quasi-convex problems (Sect. 2) and show that their convergence rate automatically adjust the favorable local structure of the optimal solution. After that, in Sects. 3 and 4, we analyze new schemes for

minimizing max-type convex functions and functions with smooth strongly convex components. Our main theoretical improvement there is the scheme with the linear rate of convergence for the problem, where the classical step-size strategy ensures only a sublinear rate.

The remaining part of the paper is devoted to problems with non-trivial functional constraints. In Sect. 5, we present a method for solving problems with single constraint in the composite form (it seems that this formulation is new). For treating this kind of problems, we introduce a new condition of *diverging series of squares* of the control step-size parameters and define *divergence delay*, which shows how quickly the method can eliminate the past. Our new method can automatically adjust to the best Hölder class containing the functional components.

In the last three sections of the paper, we present methods, which can efficiently approximate optimal Lagrange Multipliers by a simple switching strategies. All these schemes ensure the optimal rate of convergence. They differ one from another by involvement of Slater condition into the final efficiency bound and by the way they treat unbounded feasible sets.

Our new technique is based on ability to solve some auxiliary univariate problems, related to the prox-type operations. However, for the simple feasible sets, these problems are easy. As a benefit, we get new methods which can adjust to the problem structure and move with much longer steps, especially in the beginning of the process.

References

1. Juditsky, A., Nemirovski, A.: First order methods for nonsmooth convex large-scale optimization. Optim. Mach. Learn. 121–148 (2011)
2. Nemirovsky, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
3. Nesterov, Y.: Minimization methods for nonsmooth convex and quasi-convex functions. Ekon. Mat. Metody **20**(3), 519–531 (1984). (**in Russian; translated as Matekon**)
4. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 261–283 (2009)
5. Nesterov, Y.: Lectures on Convex Optimization. Springer, Berlin (2018)
6. Nesterov, Y., Shikhman, V.: Quasi-monotone subgradient methods for nonsmooth convex minimization. JOTA **165**, 917–940 (2015)
7. Polyak, B.: Introduction to Optimization. Optimization Software, New York (1987)
8. Shor, N.: On the structure of algorithms for the numerical solution of optimal planning and design problems. Ph.D. Dissertation, Cybernetics Institute, Academy of Sciences of the Ukrainian SSR, Kiev (1964). **(in Russian)**