J Optim Theory Appl (2015) 165:917–940
DOI 10.1007/s10957-014-0677-5

Quasi-monotone Subgradient Methods for Nonsmooth Convex Minimization

Yu. Nesterov · V. Shikhman

Received: 27 April 2014 / Accepted: 1 November 2014 / Published online: 15 November 2014

Abstract In this paper, we develop new subgradient methods for solving nonsmooth convex optimization problems. These methods guarantee the best possible rate of convergence for the whole sequence of test points. Our methods are applicable as efficient real-time stabilization tools for potential systems with infinite horizon. Preliminary numerical experiments confirm a high efficiency of the new schemes.

Keywords Convex optimization · Nonsmooth optimzation · Subgradient methods · Rate of convergence · Primal-dual methods

Mathematics Subject Classification 90C25 · 90C47 · 68Q25

1 Introduction

Subgradient methods for minimizing nonsmooth convex functions have a long history of development. The first methods of this type were proposed in the early 60s (see references and bibliographical comments in monographs of pioneers of this field, Shor [1] and Polyak [2]). The minimizing sequences in these schemes are formed as Euclidean projections of the consecutive shifts along the directions of subgradients

Communicated by Amir Beck.

Yu. Nesterov (✉) · V. Shikhman
Center for Operations Research and Econometrics (CORE), 34 voie du Roman Pays,
1348 Louvain-la-Neuve, Belgium
e-mail: Yurii.Nesterov@uclouvain.be

V. Shikhman
e-mail: Vladimir.Shikhman@uclouvain.be

onto the feasible set. The only degree of freedom in these schemes was related to the choice of step-size parameters.

For these methods, Euclidean framework is essential. Indeed, in contrast to differentiable functions, in the nonsmooth case, an arbitrary subgradient *cannot* serve as a descent direction for the current test point. Hence, the only reliable Lyapunov function for establishing convergence of these processes is the squared Euclidean distance from the current test point to one of the optimal solutions of the minimization problem.

Next big step in the development of subgradient schemes was done in the famous monograph by Nemirovski and Yudin [3]. It was related to clarification of several important aspects. First of all, it was proven that for Euclidean setup, the complexity estimates of the simplest subgradient methods are proportional to the uniform (in the dimension) lower complexity bound for the nonsmooth minimization problems. Thus, these methods are *optimal* in the Euclidean setup.

At the same time, it was observed that the complexity bounds heavily depend on the *size* of the feasible set. This value was traditionally defined with respect to Euclidean norm. However, the size of the same set, measured in different norms, can be very different. How it is possible to take into account a particular geometry of the feasible set?

For that, it was suggested to use special *prox-functions*, which are strongly convex on the feasible set with respect to a certain norm. Using these functions, it is possible to define a mapping from the dual space (containing subgradients) into the primal space, which contains our variables. The new *dual methods*, updating at each iteration the dual model of the objective function, were called the *Mirror Descent Methods* (MDM).

Despite to its mathematical beauty, MDM have a hidden inconsistency. Indeed, new subgradients are included in their linear model with *vanishing weights*. This contradicts to one of the basic principles of convergent iterative schemes, which tells us that during the process the importance and quality of new information should increase. This drawback was eliminated in the *Dual Averaging Methods* [4], which introduce in the process a new control sequence of *scaling coefficients*. This small modification allows the increasing weights for the new elements of dual model, keeping their rate of convergence at the optimal level.

Recently, it became clear that all methods mentioned above have a common drawback:

They cannot generate a convergent sequence of test points.

For all methods known so far, it is possible to guarantee only convergence of the *average values* of the objective function along the minimizing sequence. This allows uncontrollable jumps of the function values at some iterations. One of the ways to escape from this difficulty is to consider the sequence of the best values of the objective achieved so far. However, this cannot be done in the situations, where we are not able to compute the values of objective function (we give an example of such an application in Sect. 4.3). Another possibility would be to define the sequence of average points of the trajectory. However, this does not work when we want to use subgradient method as an adjustment strategy for approaching a stable state of some system.

The main goal of this paper consists in the development of *quasi-monotone* subgradient methods. For such primal-dual methods, we are able to justify the rate of convergence for the whole sequence of test points.

2 Comparative Analysis of Existing Subgradient Methods

Consider the problem

$$f_* := \min_{x \in Q} f(x) \tag{1}$$

with convex nondifferentiable objective function f, and closed and convex feasible set $Q \subseteq \mathbb{R}^n$, $\mathrm{dom}\, f \subseteq Q$. In [1] and [2], it was suggested to apply the simplest *Subgradient Method*

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \tag{2}$$

where $\pi_Q(x)$ is a Euclidean projection of point x onto the set Q, $\nabla f(x_k)$ is arbitrary element from subdifferential $\partial f(x_k)$,[1] and $a_k > 0$ is a step-size parameter. The rate of convergence of this method can be easily obtained from the following inequality, which is derived from analysis of variation of the Euclidean distance between the current iteration and some optimal solution x_* of problem (1):

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\tfrac{1}{2} \|x_0 - x_*\|_2^2 + \tfrac{1}{2} \sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_2^2 \right], \tag{3}$$

where $A_t = \sum\limits_{k=0}^{t} a_k$. In order to have the right-hand side of this inequality vanishing, it is enough to ensure

$$\lim_{t \to \infty} a_t = 0, \quad \lim_{t \to \infty} A_t \ = \ \infty. \tag{4}$$

Note that the first of these conditions looks reasonable since for nonsmooth functions we cannot expect subgradients to be vanishing in a neighborhood of the optimal solution. Hence, this condition is absolutely necessary for convergence of the scheme (2).

From the complexity point of view, the best rule for the choice of step-size parameters is as follows:

$$a_t = \frac{R}{L\sqrt{t+1}}, \quad t \geq 0, \tag{5}$$

[1] For the sake of notation, we assume that this subgradient is uniquely defined by the argument. At the points of nondifferentiability, this, in general, is not true. However, we assume that at such points the first-order oracle always returns the same answer. The same convention is used for all convex functions in this paper (e.g. prox-functions; see below).

where R is an upper bound for initial distance $\|x_0 - x_*\|_2$, and L is an upper bound for the norm of subgradients:

$$\|\nabla f(x)\|_2 \le L, \quad x \in Q. \tag{6}$$

In this case, an ϵ-approximation of the optimal value f_* of problem (1) can be found in

$$O\left(\frac{L^2 R^2}{\epsilon^2}\right) \tag{7}$$

iterations of method (2).

In order to describe the framework of *Mirror Descent Methods* [3], we need to introduce the special *prox-function* $d(\cdot)$, which must be strongly convex on the feasible set Q:

$$d(y) \ge d(x) + \langle \nabla d(x), y - x\rangle + \tfrac{1}{2}\|y - x\|^2, \quad x, y \in Q, \tag{8}$$

and attain its minimum on Q at some point x_0 with $d(x_0) = 0$. In definition (8), we can use already an arbitrary norm $\|\cdot\|$. For Euclidean framework, we can choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$.

Note that method (2) is essentially *primal*. It generates points directly in the feasible set Q, which is contained in the primal space of variables, say $\mathbb{E}$. At the same time, any subgradient, by its origin, defines a *linear function* on $\mathbb{E}$. Hence, it belongs to the *dual space* $\mathbb{E}^*$. The updating rule in (2) is consistent only because we identify $\mathbb{E}$ with $\mathbb{R}^n$ and consequently $\mathbb{E}^* = \mathbb{R}^n$.

Mirror Descent Method was the first *dual method*, which works directly in the dual space. At each iteration, it updates a *linear model* of the objective function and maps it back into the primal space:

$$x_{t+1} = \min_{x \in Q}\left\{ \left\langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \right\rangle + d(x)\right\}, \quad t \ge 0. \tag{9}$$

The rate of convergence of this scheme can be obtained from inequality

$$\frac{1}{A_t}\sum_{k=0}^{t} a_k f(x_k) - f_* \le \frac{1}{A_t}\left[d(x_*) + \tfrac{1}{2}\sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_*^2 \right], \tag{10}$$

which coincides with (3) up to definition of distances and norms:

$$\|s\|_* = \max_{x \in \mathbb{E}}\{\langle s, x\rangle : \ \|x\| \le 1\}, \quad s \in \mathbb{E}^*. \tag{11}$$

Thus, the convergence of scheme (9) is guaranteed by the same conditions (4), and the choice of parameters (5) results in the efficiency bound (7), where $R^2 \geq d(x_*)$ and L is defined by (6) with respect to the dual norm $\| \cdot \|_*$.[2]

However, note that in MDM, the new subgradients are included in the linear model $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$ with *vanishing weights* (see (4)). This drawback was eliminated in the *Dual Averaging Methods* [4]:

$$x_{t+1} = \min_{x \in Q} \left\{ \left\langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \right\rangle + \gamma_t d(x) \right\}, \quad t \geq 0, \tag{12}$$

which introduce in the process (9) a new control sequence of *scaling coefficients* $\{\gamma_t\}_{t \geq 0}$. This small modification results in the following estimate:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\gamma_t d(x^*) + \sum_{k=0}^{t} \frac{a_k^2}{2\gamma_k} \|\nabla f(x_k)\|_*^2 \right]. \tag{13}$$

It can be easily seen that now we have much more freedom in the choice of averaging coefficients $\{a_t\}_{t \geq 0}$. For example, we can choose $a_t = 1$ for all $t \geq 0$. Then the choice $\gamma_t = \frac{L}{R}\sqrt{t+1}$ ensures for this method the optimal complexity bound (7).

As we have already mentioned, all methods above cannot generate a convergent sequence of test points. We can guarantee only that

$$\lim_{t \to \infty} \frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) = f_*.$$

Clearly, this fact allows uncontrollable jumps of the function values at some iterations. One of the ways to escape from this difficulty is to consider the sequence of the record values

$$f_t^* = \min_{0 \leq k \leq t} f(x_k).$$

However, this cannot be done in the situations, where we are not able to compute the values of objective function (we give an example of such application in Sect. 4.3). Another possibility is to define the sequence of points

$$\bar{x}_t = \frac{1}{A_t} \sum_{k=0}^{t} a_k x_k.$$

[2] In paper [5], Beck and Teboulle justified a *primal* subgradient method, which works with Bregman distances: $x_{t+1} = \min\limits_{x \in Q} \{a_t \langle \nabla f(x_t), x\rangle + D(x_t, x)\}$, where $D(x, y) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle$. The rate of convergence of this method can be derived from the same inequality (10). In our terminology, this is a pure primal scheme since it does not maintain a linear model of the objective function.

Then, in view of convexity, we have $\lim_{t\to\infty} f(\bar{x}_t) = f_*$. However, this suggestion is not good for some applications, where we want to use a subgradient method as an adjustment strategy for approaching a stable state of some system. In this case, the variable x_t has interpretation of the current state of control parameters, and the sugradient at x_t represents the observed reaction of the system. It is important to implement the variants of control, which asymptotically stabilize the system. In such models, it is not reasonable to accumulate knowledge about potentially good variants, which will be never implemented in practice.

Contents We derive our methods from a relaxed version of the estimate sequence condition (see Sect. 2.2.1 in [6]), where we allow more freedom in the right-hand side of the recursively updated inequalities. This technique is presented in Sect. 3. Our first method for solving the problem (1) is the subgradient method with double averaging. It can be seen as an augmentation of method (12) by one additional averaging operation in the primal space, which is performed at each iteration. As a result, we can prove the rate of convergence for the whole sequence of test points. In the same section, we present a variant with triple averaging, which has slightly better performance guarantees. In both schemes we give convergence conditions for a wide range of control parameters and discuss the best strategies for their choice.

In Sect. 4 we discuss several applications, where it is possible to generate approximate primal-dual optimal solutions. We start from the convergence results for primal-dual Fenchel problem. In Sect. 4.1, it is shown how to reconstruct primal and dual solution of minimax problem with known structure. In Sect. 4.2, we demonstrate that, by solving the Lagrangian dual of the primal problem with functional constraints, we can easily approach an optimal primal solution. Finally, in Sect. 4.3 we consider a model of taxation for an industry, generating pollution. The utility functions of the producers are not known to the center. However, it can measure the generated pollution, which corresponds to the current level of taxes. We show that even in this situation, the taxation center can apply a real-time strategy, which converges in the limit to the optimal values of taxes. Moreover, during the adjustment process, the producers will be willing to apply in average the socially optimal production strategies.

In Sect. 5, we present the results of preliminary computational experiments. They demonstrate that new methods outperform the standard minimization schemes on certain problem instances. In the last Sect. 6 we discuss the obtained results and further perspectives for development of subgradient schemes.

Notation We denote by $\mathbb{E}$ a finite-dimensional linear vector space, and by $\mathbb{E}^*$ its dual space. For $x \in \mathbb{E}$ and $s \in E^*$ denote by $\langle s, x\rangle$ the value of the linear function s at x. For closed and convex function f, denote by $f^*(\cdot)$ its Fenchel conjugate:

$$f^*(s) = \sup_{x\in\mathbb{E}}[\langle s, x\rangle - f(x)], \quad s \in \mathbb{E}^*. \tag{14}$$

Since function f is closed, we have (e.g., [7])

$$f(x) = \max_{s\in\mathbb{E}^*}[\langle s, x\rangle - f^*(s)], \quad x \in \mathrm{dom}\, f. \tag{15}$$

Sometimes it is useful to define conjugate functions with respect to a set. Consider a closed function f and a closed and convex set $C \subseteq \mathbb{E}$. Denote

$$f_C^*(s) = \sup_{x \in C} [\langle s, x \rangle - f(x)], \quad s \in \mathbb{E}^*. \tag{16}$$

If $\mathbb{E} = \mathbb{R}^n$, then $\mathbb{E}^* = \mathbb{R}^n$ and $\langle s, x \rangle = \sum\limits_{i=1}^n x^{(i)} s^{(i)}$ for $x, s \in \mathbb{R}^n$. In these spaces, we use the standard notation for ℓ_p-norms with $p \geq 1$:

$$\|x\|_p = \left[\sum_{i=1}^n |x^{(i)}|^p \right]^{1/p}, \quad x \in \mathbb{R}^n.$$

Finally, $0_n \in \mathbb{R}^n$ denotes the vector of all zeros, and $1_n \in \mathbb{R}^n$ denotes the vector of all ones.

3 Methods with Multiple Averaging

In this section, we consider the following minimization problem:

$$\min_{x \in Q} f(x), \tag{17}$$

where Q is a closed and convex set in finite-dimensional linear vector space $\mathbb{E}$, and f is a closed convex function on $\mathbb{E}$, such that $Q \subseteq \operatorname{dom} f \subseteq \mathbb{E}$. We assume that the set Q is *simple* (see below).

For function $f(\cdot)$, we denote by $\nabla f(x)$ its arbitrary subgradient at $x \in Q$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad y \in Q. \tag{18}$$

Suppose that problem (17) is solvable and denote by x_* its optimal solution, $f_* := f(x_*)$. It is convenient to assume that $\operatorname{int} Q \neq \emptyset$ (otherwise we work with relative interior of Q). For the set Q, we assume to be known a *prox-function* $d(x)$, satisfying the following assumption.

Assumption 3.1 1. $d(x) \geq 0$ for all $x \in Q$ and $d(x_0) = 0$ for some $x_0 \in Q$.
2. $d(x)$ is strongly convex on Q with convexity parameter one:

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{2}\|y - x\|^2, \quad x, y \in Q. \tag{19}$$

3. Auxiliary minimization problem

$$\min_{x \in Q} [\langle s, x \rangle + \gamma d(x)], \quad s \in \mathbb{E}^*, \tag{20}$$

is easily solvable. Denote by $x_\gamma(s)$ its unique solution.

In this section we always assume that Assumption 3.1 is satisfied.

First item Assumption 3.1 guarantees that $\langle \nabla d(x_0), x - x_0 \rangle \geq 0$ for all $x \in Q$. Thus, for any $x \in Q$ we have

$$d(x) \geq d(x_0) + \langle \nabla d(x_0), x - x_0 \rangle + \tfrac{1}{2}\|x - x_0\|^2 \;\geq\; \tfrac{1}{2}\|x - x_0\|^2. \tag{21}$$

For proving the convergence of optimization methods as applied to problem (17), we use a relaxed version of the estimate sequences technique (e.g., Sect. 2.2.1 in [6]). We are going to generate a minimizing sequence $\{x_t\}_{t\geq 0} \subset Q$, satisfying the following condition:

$$A_t f(x_t) \leq \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t \quad \forall x \in Q, \tag{22}$$

where $\{a_k\}_{k\geq 0}$ and $\{\gamma_t\}_{t\geq 0}$ are sequences of positive parameters, $A_t = \sum\limits_{k=0}^{t} a_k$, and all B_t are nonnegative.

Denote $\ell_t(x) = \sum\limits_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle]$, and

$$\psi_t^* = \min_{x \in Q} [\ell_t(x) + \gamma_t d(x)].$$

Thus, condition (22) can be rewritten in the following form:

$$A_t f(x_t) \leq \psi_t^* + B_t. \tag{23}$$

Let us derive some straightforward consequences of the above condition. Denote

$$s_t = \frac{1}{A_t} \sum_{k=0}^{t} a_k \nabla f(x_k).$$

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{\langle s, x \rangle : \; x \in C\}, \quad s \in \mathbb{E}^*. \tag{24}$$

Lemma 3.1 *Let the sequence of points $\{x_t\}_{t\geq 0}$ satisfy condition (22). Then for any $t \geq 0$ we have:*

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \leq \frac{1}{A_t}(B_t + \gamma_t D_C), \tag{25}$$

where $D_C = \max\limits_x \{d(x) : \; x \in C \bigcap Q\}$.

Proof In view of condition (22), for any $x \in Q$ and $y \in \mathbb{E}$, we have

$$\begin{aligned}
&\sum_{k=0}^{t} a_k f(x_k) + \gamma_t d(x) + B_t \\
&\quad \geq A_t f(x_t) + A_t \langle s_t, y - x \rangle + \sum_{k=0}^{t} a_k \langle \nabla f(x_k), x_k - y \rangle \\
&\quad \overset{(18)}{\geq} A_t f(x_t) + A_t \langle s_t, y - x \rangle + \sum_{k=0}^{t} a_k f(x_k) - A_t f(y).
\end{aligned}$$

Thus, $\frac{1}{A_t}(B_t + \gamma_t d(x)) \geq f(x_t) + [\langle s_t, y \rangle - f(y)] + \langle -s_t, x \rangle$, and we get (25) in view of definition of D_C, (14), and (24). □

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{\langle s, x_* - x \rangle : \ \|x - x_*\| \leq R\}, \quad s \in \mathbb{E}^*. \tag{26}$$

Note that $\|s\|_R^* \geq 0$ for any $s \in \mathbb{E}^*$. On the other hand, in view of the first-order optimality condition for problem (17), there exists $g_* \in \partial f(x_*)$ such that $\langle g_*, x - x_* \rangle \geq 0$ for all $x \in Q$. Therefore $\|g_*\|_R^* = 0$. Thus, the value $\|s\|_R^*$ measures the quality of hyperplane defined by s, playing the role of separator between the feasible set Q and the level set $\{x \in \mathbb{E} : \ f(x) \leq f_*\}$.

Corollary 3.1 *Let a sequence of points $\{x_t\}_{t \geq 0}$ satisfy condition (22). Then for any $t \geq 0$ we have*

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{A_t}(B_t + \gamma_t G_R), \tag{27}$$

where $G_R = \max_{x \in Q} \{d(x) : \ \|x - x^*\| \leq R\}$.

Proof Let us choose $C = \{x \in Q : \ \|x - x_*\| \leq R\}$. Then, in view of Lemma 3.1 we have

$$\frac{1}{A_t}(B_t + \gamma_t G_R) \geq f(x_t) + \langle s_t, x_* \rangle - f_* + \langle -s_t, x \rangle, \quad x \in C.$$

Maximizing the right-hand side of this inequality in x, we obtain (27) from (26). □

Note that $G_R > G_0 = d(x_*)$.

It remains to find a recursive strategy for maintaining condition (22). Consider the following process.

Subgradient Method with Double Averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{A_t \langle s_t, x \rangle + \gamma_t d(x)\}$. (28)

2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

Thus, $x_t^+ = \arg\min\limits_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$. It is easy to see that

$$x_t = \frac{1}{A_t} \left[a_0 x_0 + \sum_{k=0}^{t-1} a_{k+1} x_k^+ \right], \quad t \geq 0. \tag{29}$$

Note that for $\tau_t \equiv 1$ method (28) coincides with the *primal-dual averaging scheme* (12). If $\tau_t \equiv 1$ and $\gamma_t \equiv 1$, then this is the *mirror descent method* (9). Additional averaging parameters in (29) make the primal sequence more stable and lead to its convergence in function value.

Theorem 3.1 *Let the sequence* $\{x_t\}_{t \geq 0}$ *be generated by method (28) with monotone sequence of parameters* $\{\gamma_t\}_{t \geq 0}$*:*

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0. \tag{30}$$

Then condition (22) holds with

$$B_t = \tfrac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \| \nabla f(x_k) \|_*^2, \tag{31}$$

where $\gamma_{-1} = \gamma_0$*. Moreover,*

$$\frac{1}{\gamma_t} A_t (f(x_t) - f_*) + \tfrac{1}{2} \| x_t^+ - x_* \|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t. \tag{32}$$

Finally, if $x_0 \in \operatorname{int} Q$*, then* $x_t \in \operatorname{int} Q$ *for all* $t \geq 0$.

Proof Indeed, assume that condition (22) is valid for some $t \geq 0$. Then

$$\begin{aligned} \psi_{t+1}^* &= \min_{x \in Q} \{ \ell_t(x) + a_{t+1} [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle] + \gamma_{t+1} d(x) \} \\ &\overset{(30)}{\geq} \min_{x \in Q} \{ \ell_t(x) + \gamma_t d(x) + a_{t+1} [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle] \} \\ &\overset{(19)}{\geq} \min_{x \in Q} \{ \psi_t^* + \tfrac{1}{2} \gamma_t \| x - x_t^+ \|^2 + a_{t+1} [f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1} \rangle] \} \end{aligned}$$

$$\begin{aligned}&\overset{(22)}{\geq} \min_{x\in Q}\{A_t f(x_t) - B_t + \tfrac{1}{2}\gamma_t\|x - x_t^+\|^2\\&\quad + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\}\\&\overset{(18)}{\geq} \min_{x\in Q}\{A_t[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1}\rangle] - B_t + \tfrac{1}{2}\gamma_t\|x - x_t^+\|^2\\&\quad + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\}.\end{aligned}$$

Since $(A_t + a_{t+1})x_{t+1} = A_t x_t + a_{t+1}x_t^+$, we obtain

$$\begin{aligned}\psi_{t+1}^* &\geq A_{t+1}f(x_{t+1}) - B_t + \min_{x\in Q}\{\tfrac{1}{2}\gamma_t\|x - x_t^+\|^2 + a_{t+1}\langle \nabla f(x_{t+1}), x - x_t^+\rangle\}\\&\geq A_{t+1}f(x_{t+1}) - B_t - \frac{a_{t+1}^2}{2\gamma_t}\|\nabla f(x_{t+1})\|_*^2 \;=\; A_{t+1}f(x_{t+1}) - B_{t+1}.\end{aligned}$$

It remains to note that

$$\begin{aligned}\psi_0^* &= \min_{x\in Q}\left\{a_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \tfrac{1}{2}\gamma_0 d(x)\right\}\\&\overset{(21)}{\geq} A_0 f(x_0) - \frac{a_0^2}{\gamma_{-1}}\|\nabla f(x_0)\|_*^2\end{aligned}$$

(recall that we define $\gamma_{-1} = \gamma_0$).

Let us prove now inequality (32). In view of Step 1 of method (28), we have

$$\begin{aligned}A_t\langle s_t, x_*\rangle + \gamma_t d(x_*) &\overset{(19)}{\geq} A_t\langle s_t, x_t^+\rangle + \gamma_t d(x_t^+) + \tfrac{1}{2}\gamma_t\|x_t^+ - x_*\|^2\\&= \psi_t^* - \sum_{k=0}^{t} a_k[f(x_k) - \langle \nabla f(x_k), x_k\rangle] + \tfrac{1}{2}\gamma_t\|x_t^+ - x_*\|^2\\&\overset{(23)}{\geq} A_t f(x_t) - B_t - A_t f_* + A_t\langle s_t, x_*\rangle + \tfrac{1}{2}\gamma_t\|x_t^+ - x_*\|^2.\end{aligned}$$

□

Corollary 3.2 *For all $t \geq 0$ we have*

$$\tfrac{1}{2}\|x_t - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma_t}B_t. \tag{33}$$

Proof In view of (29), each point x_t belongs to a convex hull of point x_0 and points $x_0^+, \ldots, x_{t-1}^+$. Hence, (33) follows from (32). □

The most important version of method (28) corresponds to $a_t = 1$, $t \geq 0$. In this case $A_t = t + 1$, and method (28) becomes dependent only on the choice of the parameters $\{\gamma_t\}_{t\geq 0}$.

Subgradient Method with Double Simple Averaging

$$\textbf{1}.\ \text{Compute } x_t^+ = \arg\min_{x\in Q}\left\{\langle\sum_{k=0}^{t}\nabla f(x_k), x\rangle + \gamma_t d(x)\right\}. \tag{34}$$

$$\textbf{2}.\ \text{Update } x_{t+1} = \tfrac{t+1}{t+2}x_t + \tfrac{1}{t+2}x_t^+.$$

For this method, we have $s_t = \frac{1}{t+1}\sum\limits_{k=0}^{t}\nabla f(x_k)$ and $x_t \overset{(29)}{=} \frac{1}{t+1}\left(x_0 + \sum\limits_{k=0}^{t-1} x_k^+\right)$.

Theorem 3.2 *Let sequence $\{x_t\}_{t\geq 0}$ be generated by method (34) with parameters $\{\gamma_t\}_{t\geq 0}$ satisfying condition (30). Then, for any $t \geq 0$, we have*

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1}\left(\gamma_t G_R + \tfrac{1}{2}\sum_{k=0}^{t}\frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}}\right). \tag{35}$$

Moreover, if $x_0 \in \operatorname{int} Q$, then $x_t \in \operatorname{int} Q$ for all $t \geq 0$.

Proof Indeed, the estimate (35) can be obtained from Theorem 3.1, taking into account representation (31) and the estimate (27). □

From now on, we assume that sugradients of function $f(\cdot)$ are uniformly bounded on int Q:

$$\|\nabla f(x)\|_* \leq L, \quad x \in \operatorname{int} Q. \tag{36}$$

Corollary 3.3 *Assume that in method (34) we have*

$$\gamma_t \to \infty, \quad \frac{\gamma_t}{t+1} \to 0. \tag{37}$$

Then $\lim\limits_{t\to\infty} f(x_t) = f_*$ and $\lim\limits_{t\to\infty}\|s_t\|_R^* = 0$.

Proof For any positive constant c, there exists a moment T such that $\gamma_t \geq c$ for all $t \geq T$. Therefore, the right-hand side of inequality (35) can be estimated from above as follows:

$$\frac{1}{t+1}\left[\gamma_t G_R + \frac{L^2}{2}\left(\sum_{k=0}^{T-1}\frac{1}{\gamma_{k-1}} + \frac{t-T}{c}\right)\right].$$

In view of conditions (37), this bound goes to $\frac{1}{c}$ as $t \to \infty$. Since c can be chosen arbitrarily large, we prove the statement. □

Let us present now the optimal strategy for choosing the values γ_t, $t \geq 0$. Consider the following sequence:

$$\gamma_t = \gamma\sqrt{t+1}, \quad t \geq 0, \tag{38}$$

where γ is a positive parameter. Note that for a convex univariate function $\xi(\tau)$, $\tau \in \mathbb{R}$, and integer bounds a, b, we have

$$\tfrac{1}{2}(\xi(a) + \xi(b)) + \int\limits_a^b \xi(\tau)d\tau \;\leq\; \sum_{k=a}^{b} \xi(k) \leq \int\limits_{a-1/2}^{b+1/2} \xi(\tau)d\tau. \tag{39}$$

Therefore, for γ_t defined by (38), we have

$$\begin{aligned}\sum_{k=0}^{t} \frac{1}{\gamma_{k-1}} \;&=\; \frac{1}{\gamma_0} + \sum_{k=0}^{t-1} \frac{1}{\gamma_k} \overset{(38)}{=} \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{k=0}^{t-1} \frac{1}{\sqrt{k+1}} \\ &\overset{(39)}{\leq} \frac{1}{\gamma} + \frac{2}{\gamma}\left(\sqrt{t+\frac{1}{2}} - \sqrt{\frac{1}{2}}\right) \leq \frac{2}{\gamma}\sqrt{t+1}.\end{aligned} \tag{40}$$

Substituting this estimate in the right-hand sides of inequalities (32) and (35), we get the following corollary.

Corollary 3.4 *Let objective function of problem (17) satisfy condition (36), and the sequence $\{\gamma_t\}_{t \geq 0}$ be defined by the rule (38). Then, for any $t \geq 0$, we have*

$$\begin{aligned} f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{\sqrt{t+1}}\left(\gamma G_R + \frac{1}{\gamma}L^2\right), \\ \frac{1}{\gamma}\sqrt{t+1}\,(f(x_t) - f_*) + \tfrac{1}{2}\|x_t^+ - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma^2}L^2. \end{aligned} \tag{41}$$

For the optimal choice $\gamma = LG_R^{-1/2}$, we get the following rate:

$$f(x_t) - f_* + \|s_t\|_R^* \leq 2LG_R^{1/2} \cdot \frac{1}{\sqrt{t+1}}. \tag{42}$$

To the best of our knowledge, method (34), (38) is the first subgradient scheme, for which the rate of convergence is justified for the whole sequence of test points.

To conclude this section, let us present a slight modification of method (28), which should exhibit a more stable behavior.

Subgradient Method with Triple Averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{A_t \langle s_t, x\rangle + \gamma_t d(x)\}$.

2. Define $\hat{x}_t = \frac{\gamma_t}{\gamma_{t+1}} x_t^+ + \left(1 - \frac{\gamma_t}{\gamma_{t+1}}\right) x_0$.

3. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1-\tau_t)x_t + \tau_t \hat{x}_t$.

(43)

Theorem 3.3 *Let sequence* $\{x_t\}_{t \geq 0}$ *be generated by method (43), and parameters* $\{\gamma_t\}_{t\geq 0}$ *satisfy condition (30). Then condition (22) holds with*

$$B_t = \tfrac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_k} \|\nabla f(x_k)\|_*^2. \tag{44}$$

Moreover,

$$\frac{1}{\gamma_t} A_t(f(x_t) - f_*) + \tfrac{1}{2}\|x_t^+ - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t. \tag{45}$$

Finally, if $x_0 \in \operatorname{int} Q$*, then* $x_t \in \operatorname{int} Q$ *for all* $t \geq 0$.

Proof The proof of this theorem is very similar to the proof of Theorem 3.1. Therefore, in our reasoning we skip some intermediate arguments.

Assume that condition (22) is valid for some $t \geq 0$. Then

$$\begin{aligned}
\psi_{t+1}^* &= \min_{x \in Q} \{\ell_t(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] + \gamma_{t+1} d(x)\} \\
&= \min_{x \in Q} \{\ell_t(x) + \gamma_t d(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] \\
&\quad + (\gamma_{t+1} - \gamma_t) d(x)\} \\
&\overset{(19)}{\geq} \min_{x \in Q} \Big\{ \psi_t^* + \tfrac{1}{2}\gamma_t \|x - x_t^+\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] \\
&\quad + \tfrac{1}{2}(\gamma_{t+1} - \gamma_t)\|x - x_0\|^2 \Big\} \\
&\geq \min_{x \in Q} \Big\{ \psi_t^* + \tfrac{1}{2}\gamma_{t+1}\|x - \hat{x}_t\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] \Big\}.
\end{aligned}$$

Now we can continue the proof in the same way as in Theorem 3.1, replacing γ_t by γ_{t+1} and x_t^+ by $\hat{x}_t$. Thus, we come to the bound

$$\psi_{t+1}^* \geq A_{t+1} f(x_{t+1}) - B_t - \frac{a_{t+1}^2}{2\gamma_{t+1}} \|\nabla f(x_{t+1})\|_*^2 \;=\; A_{t+1} f(x_{t+1}) - B_{t+1}.$$

It remains to note that

$$\begin{aligned}\psi_0^* &= \min_{x\in Q}\left\{a_0[f(x_0)+\langle\nabla f(x_0),x-x_0\rangle]+\tfrac{1}{2}\gamma_0 d(x)\right\}\\ &\overset{(21)}{\geq} A_0 f(x_0)-\frac{a_0^2}{\gamma_0}\|\nabla f(x_0)\|_*^2.\end{aligned}$$

□

Different variants of this scheme, including the choice of averaging coefficients $a_t = 1$ combined with the choice (38) of scaling coefficients, can be justified exactly in the same way as the above-mentioned variants of the scheme (28). With the optimal choice of coefficients, their rate of convergence is of the order $O\left(\frac{LG_R^{1/2}}{\sqrt{t+1}}\right)$. However, note that method (28) allows a significant flexibility in the choice of parameters. We can choose, for example,

$$a_t = t, \quad \gamma_t \;=\; t^{3/2}, \quad t \geq 1.$$

In this case, $A_t = O(t^2)$, $\frac{\gamma_t}{A_t} = O(t^{-1/2})$, and $\frac{1}{A_t}\sum\limits_{k=0}^{t}\frac{a_k^2}{\gamma_{k-1}} = O(t^{-1/2})$. Thus, in view of Theorem 3.1, this choice of coefficients also gives an optimal rate of convergence.

4 Primal-Dual Aggregating Strategies

4.1 Optimization Problem with Known Minimax Structure

Consider minimization problem (17) with partially available structure of the objective function. Namely, assume that it has the following representation:

$$f(x) = \hat{f}(x) + \max_{u\in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}, \tag{46}$$

where $\hat{f}$ is a closed convex function on Q, U is a closed convex set in $\mathbb{E}_1$, A is a linear operator from $\mathbb{E}_1$ to $\mathbb{E}^*$, and $\hat{\phi}(\cdot)$ is a closed convex function on U. Denote by $u(x)$ one of the optimal solutions of optimization problem in (46). Then, by Danskin's theorem.

$$\nabla f(x) \;:=\; \nabla\hat{f}(x) + Au(x) \in \partial f(x).$$

Let us write down the adjoint problem to (17):

$$\begin{aligned}f_* &= \min_{x\in Q}\left\{\hat{f}(x)+\max_{u\in U}[\langle Au,x\rangle-\hat{\phi}(u)]\right\}\\ &= \max_{u\in U}\left\{-\hat{\phi}(u)+\min_{x\in Q}[\langle Au,x\rangle+\hat{f}(x)]\right\}\end{aligned}$$

$$\overset{(16)}{=} -\min_{u\in U}\left\{\hat{\phi}(u) + \hat{f}_Q^*(-Au)\right\}.$$

Thus, we come to the following primal-dual problem:

$$\min_{x\in Q,\, u\in U}\{\Phi(x,u) := f(x) + \hat{\phi}(u) + \hat{f}_Q^*(-Au)\}. \tag{47}$$

The optimal value of this problem is zero.

Let us show how the optimal solution of this problem can be approximated by method (28). For simplicity, assume that set Q is bounded: $d(x) \le D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Note that

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k\rangle \\ & \quad = \hat{f}(x_k) + \langle Au_k, x_k\rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + Au_k, x - x_k\rangle \\ & \quad \le \hat{f}(x) + \langle Au_k, x\rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t}\sum\limits_{k=0}^{t} a_k u_k \in U$. Using notation of Sect. 3, we have

$$\ell_t(x) \le A_t \hat{f}(x) + A_t\langle A\bar{u}_t, x\rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \ \le \ A_t[\hat{f}(x) + \langle A_t\bar{u}_t, x\rangle - \hat{\phi}(\bar{u}_t)].$$

Therefore,

$$\begin{aligned} \psi_t^* &= \min_{x\in Q}\{\ell_t(x) + \gamma_t d(x)\} \ \le \ \min_{x\in Q}\ell_t(x) + \gamma_t D \\ &\le A_t \min_{x\in Q}[\hat{f}(x) + \langle A\bar{u}_t, x\rangle - \hat{\phi}(\bar{u}_t)] + \gamma_t D \\ &= -A_t[\hat{\phi}(\bar{u}_t) + \hat{f}_Q^*(-A\bar{u}_t)] + \gamma_t D. \end{aligned}$$

Thus, in view of inequality (23), we get

$$\Phi(x_t, \bar{u}_t) \le \frac{1}{A_t}[\gamma_t D + B_t].$$

It remains to use the estimates for the values A_t, γ_t, and B_t obtained in Sect. 3. Note that it may be difficult to solve the primal-dual problem (47) directly, since the computation of the values and subgradients of function $\hat{f}_Q^*$ can be very difficult.

4.2 Dual Lagrangian Methods

Consider the following minimization problem:

$$f^* := \min_{x\in Q}\{f_0(x) : \ f(x) \le 0_m\}, \tag{48}$$

where $Q \subset \mathbb{E}$ is a bounded closed and convex set, function f_0 is closed and convex on Q, and the vector function $f : Q \to \mathbb{R}^m$ consists of closed and convex components. Let us form the dual Lagrangian problem to (48):

$$\begin{aligned} f^* &= \min_{x \in Q} \max_{\lambda \geq 0_m} \{ f_0(x) + \langle \lambda, f(x) \rangle \} \;\geq\; \max_{\lambda \geq 0_m} \min_{x \in Q} \{ f_0(x) + \langle \lambda, f(x) \rangle \} \\ &= \max_{\lambda \geq 0_m} \left\{ \phi(\lambda) := \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle] \right\} \;:=\; f_*. \end{aligned}$$

Let us assume that the set Q and functions f_0 and f are so simple, that the value of the dual function ϕ is computable at any $\lambda \geq 0_m$. Then by Danskin theorem

$$\nabla \phi(\lambda) = f(x(\lambda)), \quad x(\lambda) \in \operatorname{Arg} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle]. \tag{49}$$

Let us solve the dual problem

$$\max_{\lambda \geq 0_m} \phi(\lambda) \tag{50}$$

by one of the schemes based on the relaxed estimate sequence condition (22). For that, we need to define a prox-function of the feasible set. Let us choose

$$d(\lambda) = \tfrac{1}{2} \| \lambda \|_2^2, \quad \lambda_0 \;=\; 0_m.$$

We can derive now the consequences of condition

$$- A_t \phi(\lambda_t) \leq - \sum_{k=0}^{t} a_k [\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle] + \gamma_t d(\lambda) + B_t, \quad \lambda \geq 0_m \tag{51}$$

(we take into account that (50) is a concave maximization problem). Note that

$$\begin{aligned} \phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle &= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle. \end{aligned}$$

Denoting now $x_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x(\lambda_k))$, we obtain

$$\begin{aligned} -A_t \phi(\lambda_t) &\overset{(51)}{\leq} - \sum_{k=0}^{t} a_k [f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle] + \gamma_t d(\lambda) + B_t \\ &\leq -A_t f_0(x_t) - A_t \langle \lambda, f(x_t) \rangle + \gamma_t d(\lambda) + B_t. \end{aligned}$$

Therefore,

$$\begin{aligned} f_0(x_t) - \phi(\lambda_t) &\le \frac{1}{A_t} B_t + \min_{\lambda \ge 0_m} \left\{ -\langle \lambda, f(x_t) \rangle + \frac{1}{A_t} \gamma_t d(\lambda) \right\} \\ &= \frac{1}{A_t} B_t - \frac{A_t}{2\gamma_t} \| (f(x_t))_+ \|_2^2. \end{aligned}$$

Thus,

$$\begin{aligned} f_0(x_t) + \frac{A_t}{2\gamma_t} \| (f(x_t))_+ \|_2^2 - f^* &\le f_0(x_t) + \frac{A_t}{2\gamma_t} \| (f(x_t))_+ \|_2^2 - \phi(\lambda_t) \\ &\le \frac{1}{A_t} B_t, \end{aligned} \tag{52}$$

and for convergence of method (28) as applied to problem (50) we need to assume boundedness of the gradient (49).

By inequality (52), we can bound the "duality gap"

$$f_0(x_t) - \phi(\lambda_t) \le \frac{1}{A_t} B_t. \tag{53}$$

Denoting now $\hat{f}_* = \min_{x \in Q} f(x)$, we also get

$$\frac{A_t}{2\gamma_t} \| (f(x_t))_+ \|_2^2 \overset{(52)}{\le} \frac{1}{A_t} B_t + f^* - \hat{f}_*. \tag{54}$$

Taking $A_t \equiv t + 1$ and choosing γ_t by (38), we obtain $O(\frac{1}{\sqrt{t}})$ upper bounds both for the duality gap and for the feasibility measure $\| (f(x_t))_+ \|_2^2$.

4.3 Privacy-Respecting Taxation

Consider the situation when a coordination center needs to bound some undesirable consequences (e.g., pollution) of commercial activity of n producers. Every producer i decides on his reasonable production volume u_i, which can be chosen from a bounded closed convex technological set $\mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1, \ldots, n$. In the absence of tax regulation, each producer justifies his choice by maximizing a concave utility function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.

If we bound the total pollution by a certain acceptable level, a reasonable social target consists in arranging the production activity in accordance to the optimal solution of the following optimization problem

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \phi_i(u_i) : \sum_{i=1}^n P_i u_i \le b, \; u_i \in \mathcal{U}_i, \; i = 1, \ldots, n \right\}. \tag{55}$$

In this problem, $b \in \mathbb{R}^m_+$ is the vector of upper limits on different kinds of pollution, and matrix P_i transforms the production activity of ith producer into the generated pollution. It is natural to assume that $0 \in \mathcal{U}_i$, $i = 1, \dots, n$, and that $b > 0$.

Since all sets $\mathcal{U}_i$ are bounded, $i = 1, \dots, n$, the problem (55) is solvable. However, it is not easy to implement its solution in practice. Indeed, the behavior of producers is usually independent and selfish. They are not going to take into account the interests of others. In order to tackle this difficulty, coordination center is going to charge the generated pollution by some taxes $p \in \mathbb{R}^m_+$. In this case, the ith producer is forced to make his choice by solving the problem

$$f_i(p) = \max_{u_i} \left[\phi_i(u_i) - \langle p, P_i u_i \rangle : \ u_i \in \mathcal{U}_i \right], \quad i = 1, \dots, n. \tag{56}$$

Denote by $u_i(p)$ one of the optimal solutions to this problem. Then

$$P_i u_i(p) \in -\partial f_i(p).$$

In this situation, the coordination center gets a possibility to reach a kind of social balance. Indeed, let us assume that it chooses the taxes as the optimal solution to the problem

$$\min_{p \geq 0} \left\{ f(p) := \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}. \tag{57}$$

This problem is dual to the optimal distribution problem (55). The gradient of the objective function in (57) is then

$$\nabla f(p) = b - v(p), \quad v(p) \ := \ \sum_{i=1}^{n} P_i u_i(p). \tag{58}$$

Note that $-\nabla f(p)$ has interpretation of the *excessive pollution* of the system. The first-order optimality condition

$$\langle \nabla f(p_*), p - p_* \rangle \geq 0, \quad \forall p \in \mathbb{R}^m_+ \tag{59}$$

implies that for positive optimal taxes the excessive pollution is vanishing. If the optimal tax is zero, then the excessive pollution is nonpositive.

The main difficulty of coordination center with solving problem (57) is related to the fact that usually the utility functions of the producers are not known. Instead, it is possible to observe only the *aggregated pollution* $v(p)$ generated by the whole industry. Let us show how the problem (57) can be solved by the subgradient method with double simple averaging.

Let us present interpretation of the objects generated by method (34) for problem (57). We treat them as the processes in discrete time. In the primal space, the method updates the taxes $p[t]$, $t \geq 0$, starting with the initial value $p[0] = p_0 = 0$. In the dual space, we update the average excessive pollution:

$$s[t] = \frac{1}{t+1}\sum_{k=0}^{t} \nabla f(p[k]) \overset{(58)}{=} b - \frac{1}{t+1}\sum_{k=0}^{t} v(p[k]).$$

In order to apply subgradient method (34), we need to choose a prox-function for $\mathbb{R}^m_+$. Let us consider

$$d(p) = \tfrac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2, \tag{60}$$

where $\varkappa_j > 0$ are some scaling coefficients. Define

$$S[t] = -(t+1)s[t] = \sum_{k=0}^{t}(v(p[k]) - b)$$

with $S[-1] = 0$. Then the adjustment process for the taxes looks as follows:

Double Simple Averaging for Taxation ($t \geq 0$)

1. Measure the total pollution volume $v(p[t])$.

2. Update the aggregate pollution $S[t] = S[t-1] + v(p[t]) - b$.

3. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \ldots, m$.

4. Define new vector of taxes $p[t+1] = \frac{t+1}{t+2}p[t] + \frac{1}{t+2}p_+[t]$.

(61)

Note that the only information reported to the tax office consists of the current pollution level $v[t] = v(p[t])$. No private information (functions ϕ_i, sets $\mathcal{U}_i$, production plans $u_i(p[t])$) is necessary for the efficient tax regulation.

Denote by $u_i[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} u_i(p[k])$, $i = 1, \ldots, n$, the *historical averages* of production plans of the producers, reacting on the dynamic tax policy (61). Let us show that they approach the optimal solution of the socially balanced coordination problem (55).

First of all, let us find an interpretation for the linear function $\ell_t(p)$. Note that

$$\begin{aligned} f(p[k])+\langle \nabla f(p[k]), p-p[k]\rangle &= \langle b, p[k]\rangle + \sum_{i=1}^{n} [\phi_i(u_i(p[k])) - \langle p[k], P_i u_i(p_k)\rangle] \\ &\quad + \langle b - \sum_{i=1}^{n} P_i u_i(p[k]), p - p[k]\rangle \\ &= \sum_{i=1}^{n} \phi_i(u_i(p[k])) + \langle b - v[k], p\rangle. \end{aligned}$$

Therefore,

$$\begin{aligned} \psi_t^* &= \min_{p\geq 0}\{\ell_t(p) + \gamma_t d(p)\} \\ &= \min_{p\geq 0}\left\{ \sum_{k=0}^{t}\left[\sum_{i=1}^{n}\phi_i(u_i(p[k])) + \langle b - v[k], p\rangle\right] + \gamma_t d(p)\right\} \\ &\leq (t+1)\sum_{i=1}^{n}\phi_i(u_i[t]) + \min_{p\geq 0}\{-\langle S[t], p\rangle + \gamma_t d(p)\} \\ &= (t+1)\sum_{i=1}^{n}\phi_i(u_i[t]) - \frac{(t+1)^2}{\gamma_t}\sum_{j=1}^{m}\frac{\varkappa_j}{2}\left(-s^{(j)}[t]\right)_+^2. \end{aligned}$$

Thus, in view of inequality (23), we have

$$f(p[t]) - \sum_{i=1}^{n}\phi_i(u_i[t]) + \frac{t+1}{\gamma_t}\sum_{j=1}^{m}\frac{\varkappa_j}{2}\left(-s^{(j)}[t]\right)_+^2 \leq \frac{1}{t+1}B_t. \qquad (62)$$

The left-hand side of this inequality is composed by the objective function of the dual problem (57), computed at the last variant of taxes $p[t]$, objective function of the primal problem (55), computed at historical averages $\{u_i[t]\}_{i=1}^n$, and the quadratic penalty for violation of the linear inequality constraints by the historical averages.

If we choose $\gamma_t = O(\sqrt{t})$, then the coefficient of the quadratic penalty $\frac{t+1}{\gamma_t}$ will go to infinity, and the right-hand side of inequality (62) will go to zero. Therefore, we come to the following conclusion.

Theorem 4.1 *Let taxation algorithm (61) be applied with $\gamma_t = O(\sqrt{t})$. Then the taxes $p[t]$ converge to the optimal solution of problem (57). At the same time, historical averages of individual production volumes $u_i[t]$, $i = 1 \ldots n$, converge to the socially optimal solution of problem (55).*

Of course, this conclusion is valid under the condition that all producers are able to measure undesirable effects $P_i u_i$ of their activity, and that they are honest in paying taxes.

5 Numerical Experiments

Let us compare numerical performance of different subgradient schemes on one difficult nonsmooth minimization problem. Denote

$$f(x) = \max \left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}. \tag{63}$$

This is a homogeneous convex function of degree one. Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$. Consider the point $\bar{x} \in \mathbb{R}^n$ with coordinates

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

It is easy to see that $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1, \dots, n$. Therefore

$$f(\bar{x}) = f(1_n) = 1.$$

Thus, the condition number of the level sets of this function with respect to infinity norm $\kappa_\infty(f)$ is very big:

$$\kappa_\infty(f) \ge 2^{n+1} - 1. \tag{64}$$

In other words, this function is highly degenerate even for a moderate dimension and we can expect that it should be difficult for subgradient methods.

Let us choose $x_0 = 1_n$. Then $R := \|x_0 - x_*\|_2 = \sqrt{n}$ and

$$\|\nabla f(x)\|_* \le L := \sqrt{5}, \quad x \in \mathbb{R}^n.$$

We assume that the exact values of R and L are available for numerical methods.

In our numerical experiments, we compare the simplest primal gradient method PGM[3] (see (2)):

$$x_{t+1} = x_t - \frac{R}{L\sqrt{t+1}} \nabla f(x_t), \quad t \ge 0,$$

the method of simple dual averaging SDA (see (12)):

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \frac{L\sqrt{t+1}}{2R} \|x - x_0\|_2^2 \right\},$$

and the method of simple double averaging SA$_2$ (see (34)) with $\gamma_t = \frac{L}{R}\sqrt{t+1}$ and Euclidean prox-function $d(x) = \frac{1}{2}\|x - x_0\|_2^2$.

[3] Recall that, for Euclidean framework with $Q \equiv \mathbb{E}$, PGM coincides with MDM (9).

Table 1 Computational results for function (63)

Dimension	PGM	SDA	SA_2	SA_2 %	L^2R^2/ϵ^2
10	51,204	9,254	586	0.29	204,800
20	102,405	65,536	1,587	0.39	409,600
40	204,805	131,072	4,094	0.50	819,200
80	409,616	262,144	6,655	0.41	1,638,400
160	819,209	524,288	16,484	0.50	3,276,800
320	1,638,409	1,048,576	35,184	0.54	6,553,600
640	3,276,807	2,097,152	73,390	0.56	13,107,200
1,280	6,553,612	4,194,304	143,475	0.55	26,214,400
2,560	13,107,205	8,388,608	309,681	0.59	52,428,800
5,120	26,214,405	16,777,216	579,893	0.55	104,857,600
10,240	52,428,810	33,554,432	1,181,849	0.56	209,715,200

Computational results of our experiments for dimension $n = 10, \ldots, 10240$ are given in the following table. All problems were solved up to accuracy $\epsilon = 2^{-6} = 0.0156$ in the function value (the optimal value of the objective was used in the stopping criterion). The first column of the table shows the dimension of the problem. Next three columns show the number of iterations of PGM, SDA, and SA_2. Next column shows the percentage of the number of iterations of SA_2 with respect to theoretical prediction, which is shown in the last column (Table 1).

As we can see, our new scheme is a clear winner of this competition.

6 Conclusions

In this paper, we presented two new methods (28) and (43), which solve convex nonsmooth minimization problems by generating convergent minimizing sequences. To the best of our knowledge, for the existing nonsmooth optimization methods, this feature is unique. Both schemes were derived from a relaxed version of the estimating sequences condition (22), which was used before only for analyzing the fast gradient methods (e.g., Chap. 2 in [6]). We have shown that the new methods can be used for generating primal-dual solutions. Preliminary numerical experiments confirm high efficiency of the new schemes.

Acknowledgments The authors would like to thank two anonymous referees for careful reading and useful comments. Research results presented in the paper have been supported by a Grant “Action de recherche concertè ARC 04/09-315” from the “Direction de la recherche scientifique - Communautè française de Belgique” and by PremoLab, MIPT (RF gov. Grant, ag.11.G34.31.0073), with RFBR research Projects 13-01-12007 ofi_m, 14-01-00722-a.

References

1. Shor, N.Z.: Minimization Methods for Non-differentiable Functions. Springer, Berlin (1985)
2. Polyak, B.T.: Introduction to Optimization. Software Inc., New York (1987)

3. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, New York (1983)
4. Nesterov, Yu.: Primal-dual subgradient methods for convex problems. Math. Program. **120**, 261–283 (2009)
5. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**, 167–175 (2003)
6. Nesterov, Yu.: Introductory Lectures on Convex Optimization. Kluwer, Boston (2004)
7. Rockafellar, R.T.: Convex Analisys. Princeton University Press, Princeton (1970)