2014/5

Convergent subgradient methods for nonsmooth convex minimization

Yu. Nesterov and Vladimir Shikhman

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium
http://www.uclouvain.be/core

CORE DISCUSSION PAPER
2014/5

Convergent subgradient methods for nonsmooth convex minimization

Yu. NESTEROV [1] and Vladimir SHIKHMAN[2]

January 2014

Abstract

In this paper, we develop new subgradient methods for solving nonsmooth convex optimization problems. These methods are the first ones, for which the whole sequence of test points is endowed with the worst-case performance guarantees. The new methods are derived from a relaxed estimating sequences condition, which allows reconstruction of the approximate primal-dual optimal solutions.

Our methods are applicable as efficient real-time stabilization tools for potential systems with infinite horizon. As an example, we consider a model of privacy-respecting taxation, where the center has no information on the utility functions of the agents. Nevertheless, we show that by a proper taxation policy, the agents can be forced to apply in average the socially optimal strategies.

Preliminary numerical experiments confirm a high efficiency of the new methods.

Keywords: convex optimization, nonsmooth optimization, subgradient methods, rate of convergence, primal-dual methods, privacy-respecting tax policy.

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: yurii.nesterov@uclouvain.be
[2] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: Vladimir.shikhman@uclouvain.be

The research results presented in the paper have been supported by a grant 'Action de recherche concertée ARC 04/09-315' from the 'Direction de la recherche scientifique' – Communauté française de Belgique' and by PremoLab, MIPT (RF gov. grant, ag. 11.G34.31.0073) with RFBR res.projects 13-01-12007 ofi-m, 14-01-00722-a. It also presents research results of the P7/36 PAI project COMEX, part of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

1 Introduction

Motivation. Subgradient methods for minimizing nonsmooth convex functions have already sufficiently long history of developments. First methods of this type were proposed in the early 60s (see references and bibliographical comments in monographs of the pioneers of this field, N. Shor [8] and B. Polyak [7]). For solving the problem

$$f_* \;=\; \min_{x \in Q} f(x) \tag{1.1}$$

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

$$x_{t+1} \;=\; \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \tag{1.2}$$

where $\pi_Q(x)$ is a Eucleaden projection of point x onto the set Q, $\nabla f(x_k)$ is arbitrary element from subdifferential $\partial f(x_k)$, and $a_k > 0$ is a step size parameter. Euclidean framework was essential for discovering this scheme. Indeed, in the contrast to the differentiable functions, in 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 the process (1.2) is the squared Euclidean distance from the current test point to one of the optimal solutions x_* to problem (1.1).

Having this picture in mind, it is easy to analyze the convergence of this scheme, taking into account the following inequality:

$$\tfrac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \;\leq\; \tfrac{1}{A_t} \left[\tfrac{1}{2} \| x_0 - x_* \|_2^2 + \tfrac{1}{2} \sum_{k=0}^{t} a_k^2 \| \nabla f(x_k) \|_2^2 \right], \tag{1.3}$$

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

$$\lim_{t \to \infty} a_t \;=\; 0, \quad \lim_{t \to \infty} A_t \;=\; \infty. \tag{1.4}$$

Note that the first of these conditions looks reasonable since for nonsmooth functuions we cannot expect subgradients be vanishing in a neighborhood of the optimal solution. Hence, this condition is absolutely necessary for convergence of the scheme (1.2).

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

$$a_t \;=\; \tfrac{R}{L\sqrt{t+1}}, \quad t \geq 0, \tag{1.5}$$

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

$$\| \nabla f(x) \|_2 \;\leq\; L, \quad x \in Q. \tag{1.6}$$

In this case, an ϵ-approximation of the optimal value f_* of problem (1.1) can be found in

$$O\left(\tfrac{L^2 R^2}{\epsilon^2} \right) \tag{1.7}$$

iterations of method (1.2).

The next big step in the development of subgradient schemes was done in the famous monograph by A. Nemirovski and D. Yudin [3]. It is related to clarification of several important aspects. First of all, it was proven that for Euclidean setup, the complexity estimate (1.7) is proportional to the uniform lower complexity bound of problem (1.1), which is valid for all dimensions of the space of variables. In this sense, scheme (1.2) is an *optimal method* for solving problem (1.1) in Euclidean setup.

At the same time, it was observed that the complexity bound (1.7) heavily depends on the *size parameter* R. Its value is defined there 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 properly into account geometry of a particular feasible set? For that, it was suggested to use a special *prox-function* $d(\cdot)$, which must be strongly convex on the feasible set Q:

$$d(y) \quad \geq \quad d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{2}\|y - x\|^2, \quad x, y \in Q, \tag{1.8}$$

and attain its minimum on Q at some point x_0 with $d(x_0) = 0$. In definition (1.8), we can use already an arbitrary norm $\|\cdot\|$. In order to incorporate this function into minimization process, in [3] there was developed a *mirror descent scheme.*

Note that method (1.2) 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 E. Hence, it belongs to the *dual space* $\mathbb{E}^*$. The updating rule in (1.2) 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:

$$x_{t+1} \quad = \quad \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0. \tag{1.9}$$

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

$$\tfrac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \quad \leq \quad \tfrac{1}{A_t} \left[d(x^*) + \tfrac{1}{2} \sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_*^2 \right], \tag{1.10}$$

which coincides with (1.3) up to the definition of the distances:

$$\|s\|_* \quad = \quad \max_{x \in \mathbb{E}} \{ \langle s, x \rangle : \; \|x\| \leq 1 \}, \quad s \in \mathbb{E}^*. \tag{1.11}$$

Thus, the convergence of scheme (1.9) is guaranteed by the same conditions (1.4), and the choice of parameters (1.5) results in the efficiency bound (1.7), where $R^2 \geq d(x_*)$ and L is computed by (1.6) with respect to the dual norm.[1]

[1] Several years ago, A. Beck and M. Teboulle [1] 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 (1.10). In our terminology, this is a pure primal scheme since it does not maintain a linear model of the objective function.

Despite to its mathematical beauty, Mirror Descent Method (1.9) has hidden inconsistency. Indeed, new subgradients are included in the linear model $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$ with *vanishing weights* (see (1.4)). This contradicts to one of the basic principles of the 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* [5]:

$$x_{t+1} \;=\; \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0, \tag{1.12}$$

which introduce in the process (1.9) a new control sequence of *scaling coefficients* $\{\gamma_t\}_{t\geq 0}$. This small modification resulted in the following estimate:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \;\leq\; \frac{1}{A_t} \left[\gamma_t d(x^*) + \sum_{k=0}^{t} \frac{a_k^2}{2\gamma_k} \| \nabla f(x_k) \|_*^2 \right], \tag{1.13}$$

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 (1.7).

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

They cannot generate a convergent sequence of test points.

Indeed, for all methods we can guarantee only that $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$. 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 Section 3.3). Another possibility is to define the sequence of points

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

Then, in view of convexity, we have $\lim\limits_{t\to\infty} f(\bar{x}_t) = f_*$. However, this suggestion is not good for some applications, were 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 assymptotically stabilize the system. In such models, it is not very useful to accumulate some knowledge about potentially good variants, which will be never implemented in practice.

The main goal of this paper is the development of *convergent* subgradient methods. For such methods, we are able to justify the rate of convergence for the whole sequence of test points, which are the only points where we compute subgradients.

Contents. We derive our methods from a relaxed version of the estimate sequence condition (see Section 2.2.1 in [4]), where we allow more freedom in the right-hand side of the recursively updated inequalities. This technique is presented in Section 2. Our first method for solving the problem (1.1) is the subgradient method with double averaging. It can be seen as an augmentation of method (1.12) 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 Section 3 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 Section 3.1, it is shown how to reconstruct primal and dual solution of minimax problem with known structure. In Section 3.2, we demonstrate that by solving the Lagrangean dual of the primal problem with functional constraints we can easily approach an optimal primal solution. Finally, in Section 3.3 we consider a model of taxation of an industry, generating pollution. The utility functions of the producers are not known to the center. However, it can detect 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 the last Section 4, we present the results of preliminary computational experiments. They demonstrate that new methods outperform the standard minimization schemes on certain problem instances.

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 function f, denote by $f^*(\cdot)$ its Fenchel conjugate:

$$f^*(s) \quad = \quad \sup_{x \in \mathbb{E}} [\langle s, x \rangle - f(x)], \quad s \in \mathbb{E}^*. \tag{1.14}$$

Since function f is closed, we have (e.g. [2])

$$f(x) \quad = \quad \max_{s \in \mathbb{E}^*} [\langle s, x \rangle - f^*(s)], \quad x \in \operatorname{dom} f. \tag{1.15}$$

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

$$f^*_C(s) \quad = \quad \sup_{x \in C} [\langle s, x \rangle - f(x)], \quad s \in \mathbb{E}^*. \tag{1.16}$$

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 $p \geq 1$:

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

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.

2 Methods with multiple averaging

In this section we consider the following minimization problem:

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

where Q is a closed 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 $x \in Q$:

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle, \quad y \in Q. \tag{2.2}$$

Suppose that problem (2.1) is solvable and denote by x_* its optimal solution, $f_* \stackrel{\mathrm{def}}{=} f(x_*)$. 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 1 • *$d(x) \geq 0$ for all $x \in Q$ and $d(x_0) = 0$ for certain $x_0 \in Q$.*

- *$d(x)$ is strongly convex on Q with convexity parameter one:*

$$d(y) \quad \geq \quad d(x) + \langle \nabla d(x), y - x \rangle + \tfrac{1}{2}\|y - x\|^2, \quad x, y \in Q. \tag{2.3}$$

- *Auxiliary minimization problem*

$$\min_{x \in Q} [\langle s, x \rangle + \gamma d(x)], \quad s \in \mathbb{E}^*, \tag{2.4}$$

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

In this section we always assume that Assumption 1 is satisfied.

Thus, for any $x \in Q$ we have

$$d(x) \quad \geq \quad d(x_0) + \langle \nabla d(x_0), x - x_0 \rangle + \tfrac{1}{2}\|x - x_0\|^2 \;\geq\; \tfrac{1}{2}\|x - x_0\|^2. \tag{2.5}$$

For proving the convergence of optimization methods as applied to problem (2.1), we use a relaxed version of the estimate sequences technique (e.g. Section 2.2.1 in [4]). We are going to generate a minimizing sequence $\{x_t\}_{t \geq 0} \subset Q$, satisfying the following condition:

$$A_t f(x_t) \quad \leq \quad \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t \quad \forall x \in Q, \tag{2.6}$$

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

Denote $\ell_t(x) = \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle]$, and $\psi_t^* = \min_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$. Thus, condition (2.6) can be rewritten in the following form:

$$A_t f(x_t) \quad \leq \quad \psi_t^* + B_t. \tag{2.7}$$

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

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

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

$$\xi_C(s) \quad = \quad \max_x \{\langle s, x\rangle : \ x \in C\}, \quad s \in \mathbb{E}^*. \tag{2.8}$$

Lemma 1 *Let the sequence of points $\{x_t\}_{t\geq 0}$ satisfy condition (2.6). Then for any $t \geq 0$ we have:*

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \quad \leq \quad \frac{1}{A_t}(B_t + \gamma_t D_C), \tag{2.9}$$

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

Proof:
In view of condition (2.6), for any $x \in Q$ and $y \in \mathbb{E}$, we have

$$\begin{aligned}\sum_{k=0}^{t} a_k f(x_k) + \gamma_t d(x) + B_t \quad &\geq \quad A_t f(x_t) + A_t\langle s_t, y - x\rangle + \sum_{k=0}^{t} a_k\langle \nabla f(x_k), x_k - y\rangle \\ &\overset{(2.2)}{\geq} \quad A_t f(x_t) + A_t\langle s_t, y - x\rangle + \sum_{k=0}^{t} a_k f(x_k) - A_t f(y).\end{aligned}$$

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 (2.9) in view of definition of D_C, (1.14), and (2.8). □

For arbitrary $R > 0$, denote

$$\|s\|^*_R \quad = \quad \max_{x \in Q}\{\langle s, x_* - x\rangle : \ \|x - x_*\| \leq R\}, \quad s \in \mathbb{E}^*. \tag{2.10}$$

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 (2.1), 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 1 *Let the sequence of points $\{x_t\}_{t\geq 0}$ satisfy condition (2.6). Then for any $t \geq 0$ we have:*

$$f(x_t) - f_* + \|s_t\|^*_R \quad \leq \quad \frac{1}{A_t}(B_t + \gamma_t G_R), \tag{2.11}$$

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

Proof:
Let us choose $C = \{x \in Q : \ \|x - x_*\| \leq R\}$. Then, in view of Lemma 1 we have

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

Maximizing the right-hand side of this inequality in x, we obtain (2.11) from (2.10). □

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

It remains to find a recursive strategy for maintaining condition (2.6). Consider the following process.

Subgradient Method with Double Averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{A_t\langle s_t, x\rangle + \gamma_t d(x)\}$. (2.12)

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

$$x_t \quad = \quad \frac{1}{A_t}\left[a_0 x_0 + \sum_{k=0}^{t-1} a_{k+1} x_k^+\right], \quad t \ge 0. \tag{2.13}$$

Note that for $\tau_t \equiv 1$ method (2.12) coincides with the *primal-dual averaging scheme* (1.12). If $\tau_t \equiv 1$ and $\gamma_t \equiv 1$, then this is the *mirror descent method* (1.9). Additional averaging parameters in (2.13) make the primal sequence more stable and lead to its convergence in function value.

Theorem 1 *Let the sequence $\{x_t\}_{t\ge 0}$ be generated by method (2.12) with monotone sequence of parameters $\{\gamma_t\}_{t \ge 0}$:*

$$\gamma_{t+1} \quad \ge \quad \gamma_t, \quad t \ge 0. \tag{2.14}$$

Then condition (2.6) holds with

$$B_t \quad = \quad \frac{1}{2}\sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2, \tag{2.15}$$

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

$$\frac{1}{\gamma_t} A_t(f(x_t) - f_*) + \frac{1}{2}\|x_t^+ - x_*\|^2 \quad \le \quad d(x_*) + \frac{1}{\gamma_t} B_t. \tag{2.16}$$

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

Proof:

Indeed, assume that condition (2.6) is valid for some $t \geq 0$. Then

$$\begin{aligned}
\psi_{t+1}^* &= \min_{x \in Q} \{\ell_t(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] + \gamma_{t+1} d(x)\} \\
&\overset{(2.14)}{\geq} \min_{x \in Q} \{\ell_t(x) + \gamma_t d(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\} \\
&\overset{(2.3)}{\geq} \min_{x \in Q} \{\psi_t^* + \tfrac{1}{2}\gamma_t \|x - x_t^+\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\} \\
&\overset{(2.6)}{\geq} \min_{x \in Q} \{A_t f(x_t) - B_t + \tfrac{1}{2}\gamma_t \|x - x_t^+\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\} \\
&\overset{(2.2)}{\geq} \min_{x \in Q} \{A_t[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x_t - x_{t+1}\rangle] - B_t + \tfrac{1}{2}\gamma_t \|x - x_t^+\|^2 \\
&\quad + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\}.
\end{aligned}$$

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

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

It remains to note that

$$\psi_0^* \;=\; \min_{x \in Q} \left\{ a_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \tfrac{1}{2}\gamma_0 d(x) \right\} \overset{(2.5)}{\geq} A_0 f(x_0) - \tfrac{a_0^2}{\gamma_{-1}} \|\nabla f(x_0)\|_*^2$$

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

Let us prove now inequality (2.16). In view of Step 1 of method (2.12), we have

$$\begin{aligned}
A_t \langle s_t, x_* \rangle + \gamma_t d(x_*) &\overset{(2.3)}{\geq} A_t \langle s_t, x_t^+ \rangle + \gamma_t d(x_t^+) + \tfrac{1}{2}\gamma_t \|x_t^+ - x_*\|^2 \\
&= \psi_t^* - \sum_{k=0}^{t} a_k [f(x_k) - \langle \nabla f(x_k), x_k \rangle] + \tfrac{1}{2}\gamma_t \|x_t^+ - x_*\|^2 \\
&\overset{(2.7)}{\geq} A_t f(x_t) - B_t - A_t f_* + A_t \langle s_t, x_* \rangle + \tfrac{1}{2}\gamma_t \|x_t^+ - x_*\|^2.
\end{aligned}$$

□

Corollary 2 *For all $t \geq 0$ we have*

$$\tfrac{1}{2}\|x_t - x_*\|^2 \;\leq\; d(x_*) + \tfrac{1}{\gamma_t} B_t. \qquad (2.17)$$

Proof:
In view of (2.13), each point x_t belongs to a convex hull of point x_0 and points $x_0^+, \dots, x_{t-1}^+$. Hence, (2.17) follows from (2.16). $\Box$

The most important version of method (2.12) corresponds to the choice $a_t = 1$, $t \geq 0$. In this case $A_t = t+1$, and method (2.12) becomes dependent only on the choice of the parameters $\{\gamma_t\}_{t\geq 0}$.

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min\limits_{x\in Q} \left\{ \langle \sum\limits_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$. (2.18)

2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

For this method, we have $s_t = \frac{1}{t+1} \sum\limits_{k=0}^{t} \nabla f(x_k)$ and $x_t \overset{(2.13)}{=} \frac{1}{t+1} \left(x_0 + \sum\limits_{k=0}^{t-1} x_k^+ \right)$.

Theorem 2 *Let sequence $\{x_t\}_{t\geq 0}$ be generated by method (2.18) with parameters $\{\gamma_t\}_{t\geq 0}$ satisfying condition (2.14). Then, for any $t \geq 0$, we have*

$$f(x_t) - f_* + \|s_t\|_R^* \quad \leq \quad \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k)\|_*^2}{\gamma_{k-1}} \right). \tag{2.19}$$

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

Proof:
Indeed, the estimate (2.19) can be obtained from Theorem 1, taking into account representation (2.15) and the estimate (2.11). $\Box$

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

$$\|\nabla f(x)\|_* \quad \leq \quad L, \qquad x \in \operatorname{int} Q. \tag{2.20}$$

Corollary 3 *Assume that in method (2.18) we have*

$$\gamma_t \quad \to \quad \infty, \quad \frac{\gamma_t}{t+1} \to 0. \tag{2.21}$$

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 exist a moment T such that $\gamma_t \geq c$ for all $t \geq T$. Therefore, the right-hand side of inequality (2.19) 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 (2.21), 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:

$$\gamma_t \;=\; \gamma\sqrt{t+1}, \quad t \geq 0, \tag{2.22}$$

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

$$\tfrac{1}{2}(f(a)+f(b)) + \int\limits_a^b \xi(\tau)d\tau \;\leq\; \sum_{k=a}^{b} \xi(k) \;\leq\; \int\limits_{a-1/2}^{b+1/2} \xi(\tau)d\tau. \tag{2.23}$$

Therefore, for γ_t defined by (2.22), we have

$$\begin{aligned} \sum_{k=0}^{t} \frac{1}{\gamma_{k-1}} \;&=\; \frac{1}{\gamma_0} + \sum_{k=0}^{t-1} \frac{1}{\gamma_k} \overset{(2.22)}{=} \frac{1}{\gamma} + \frac{1}{\gamma}\sum_{k=0}^{t-1} \frac{1}{\sqrt{k+1}} \\ &\overset{(2.23)}{\leq} \frac{1}{\gamma} + \frac{2}{\gamma}\left(\sqrt{t+\tfrac{1}{2}} - \sqrt{\tfrac{1}{2}}\right) \;\leq\; \frac{2}{\gamma}\sqrt{t+1}. \end{aligned} \tag{2.24}$$

Substituting this estimate in the right-hand side of inequality (2.19), we get the following corollary.

Corollary 4 *Let objective function of problem (2.1) satisfy condition (2.20), and the sequence* $\{\gamma_t\}_{t\geq 0}$ *be defined by the rule (2.22). Then, for any* $t \geq 0$, *we have*

$$\begin{aligned} f(x_t) - f_* + \|s_t\|_R^* \;&\leq\; \frac{1}{\sqrt{t+1}}\left(\gamma G_R + \tfrac{1}{\gamma}L^2\right), \\ \tfrac{1}{\gamma}\sqrt{t+1}\,(f(x_t) - f_*) + \tfrac{1}{2}\|x_t^+ - x_*\|^2 \;&\leq\; d(x_*) + \tfrac{1}{\gamma^2}L^2. \end{aligned} \tag{2.25}$$

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

$$f(x_t) - f_* + \|s_t\|_R^* \;\leq\; 2LG_R^{1/2} \cdot \frac{1}{\sqrt{t+1}}. \tag{2.26}$$

To the best of our knowledge, method (2.18), (2.22) 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 (2.12), which should exhibit a more stable behavior.

Subgradient Method with Triple Averaging

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

2. Define $\hat{x}_t = \frac{\gamma_t}{\gamma_{t+1}} x_t^+ + \left(1 - \frac{\gamma_t}{\gamma_{t+1}}\right) x_0$.        (2.27)

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

Theorem 3 *Let sequence $\{x_t\}_{t \geq 0}$ be generated by method (2.27), and parameters $\{\gamma_t\}_{t \geq 0}$ satisfy condition (2.14). Then condition (2.6) holds with*

$$B_t \;=\; \tfrac{1}{2} \sum_{k=0}^{t} \tfrac{a_k^2}{\gamma_k} \|\nabla f(x_k)\|_*^2. \tag{2.28}$$

Moreover,

$$\tfrac{1}{\gamma_t} A_t (f(x_t) - f_*) + \tfrac{1}{2}\|x_t^+ - x_*\|^2 \;\leq\; d(x_*) + \tfrac{1}{\gamma_t} B_t. \tag{2.29}$$

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 1. Therefore, in our reasoning we skip some intermediate arguments.

Assume that condition (2.6) is valid for some $t \geq 0$. Then

$$\begin{array}{rcl}
\psi_{t+1}^* & = & \min\limits_{x \in Q} \{\ell_t(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] + \gamma_{t+1} d(x)\} \\[2ex]
& = & \min\limits_{x \in Q} \{\ell_t(x) + \gamma_t d(x) + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] + (\gamma_{t+1} - \gamma_t) d(x)\} \\[2ex]
& \overset{(2.3)}{\geq} & \min\limits_{x \in Q} \{\psi_t^* + \frac{1}{2}\gamma_t \|x - x_t^+\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle] \\[2ex]
& & + \frac{1}{2}(\gamma_{t+1} - \gamma_t)\|x - x_0\|^2\} \\[2ex]
& \geq & \min\limits_{x \in Q} \{\psi_t^* + \frac{1}{2}\gamma_{t+1} \|x - \hat{x}_t\|^2 + a_{t+1}[f(x_{t+1}) + \langle \nabla f(x_{t+1}), x - x_{t+1}\rangle]\}.
\end{array}$$

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

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

It remains to note that

$$\psi_0^* \;=\; \min_{x \in Q} \left\{ a_0[f(x_0) + \langle \nabla f(x_0), x - x_0\rangle] + \tfrac{1}{2}\gamma_0 d(x) \right\} \overset{(2.5)}{\geq} A_0 f(x_0) - \tfrac{a_0^2}{\gamma_0} \|\nabla f(x_0)\|_*^2.$$

□

Different variants of this scheme, including the choice of averaging coefficients $a_t = 1$ combined with the choice (2.22) of scaling coefficients, can be justified exactly in the same way as the above mentioned variants of the scheme (2.12). 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 (2.12) 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 1, this choice of coefficients also gives an optimal rate of convergence.

3 Primal-dual aggregating strategies

3.1 Optimization problem with known minimax structure

Consider minimization problem (2.1) with partially available structure of the objective function. Namely, assume that it has the following representation:

$$f(x) \quad = \quad \hat{f}(x) + \max_{u \in U} \{ \langle Au, x \rangle - \hat{\phi}(u) \}, \tag{3.1}$$

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 (3.1). Then, by Danskin's theorem.

$$\nabla f(x) \stackrel{\mathrm{def}}{=} \nabla \hat{f}(x) + Au(x) \quad \in \quad \partial f(x).$$

Let us write down the adjoint problem to (2.1):

$$\begin{aligned} f_* \quad &= \quad \min_{x \in Q} \left\{ \hat{f}(x) + \max_{u \in U} [\langle Au, x \rangle - \hat{\phi}(u)] \right\} \\ &= \quad \max_{u \in U} \left\{ -\hat{\phi}(u) + \min_{x \in Q} [\langle Au, x \rangle + \hat{f}(x)] \right\} \\ &\stackrel{(1.16)}{=} \quad -\min_{u \in U} \left\{ \hat{\phi}(u) + \hat{f}^*_Q(-Au) \right\}. \end{aligned}$$

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

$$\min_{x \in Q,\, u \in U} \{ \Phi(x,u) \stackrel{\mathrm{def}}{=} f(x) + \hat{\phi}(u) + \hat{f}^*_Q(-Au) \}. \tag{3.2}$$

The optimal value of this problem is zero.

Let us show how the optimal solution of this problem can be approximated by method (2.12). For simplicity, assume that set Q is bounded: $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Note that

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

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Using notation of Section 2, we have

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

Therefore,

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

Thus, in view of inequality (2.7), we get

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

It remains to use the estimates for the values A_t, γ_t, and B_t obtained in Section 2. Note that it may be difficult to solve the primal-dual problem (3.2) directly, since the computation of the values and subgradients of function $\hat{f}_Q^*$ can be very difficult.

3.2 Dual Lagrangian methods

Consider the following minimization problem:

$$f^* \stackrel{\mathrm{def}}{=} \min_{x \in Q} \{ f_0(x) : \ f(x) \leq 0_m \}, \tag{3.3}$$

where $Q \subset \mathbb{E}$ is a bounded closed 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 (3.3):

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

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

$$\nabla \phi(\lambda) = f(x(\lambda)), \quad x(\lambda) \in \mathrm{Arg} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle]. \tag{3.4}$$

Let us solve the dual problem

$$\max_{\lambda \geq 0_m} \phi(\lambda) \tag{3.5}$$

by one of the schemes based on the relaxed estimate sequence condition (2.6). 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

$$-A_t \phi(\lambda_t) \leq -\sum_{k=0}^{t} a_k [\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle] + \gamma_t d(\lambda) + B_t, \quad \lambda \geq 0_m \tag{3.6}$$

(we take into account that (3.5) 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

$$\begin{array}{rcl} -A_t \phi(\lambda_t) & \overset{(3.6)}{\leq} & -\sum\limits_{k=0}^{t} a_k [f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle] + \gamma_t d(\lambda) + B_t \\ & \leq & -A_t f_0(x_t) - A_t \langle \lambda, f(x_t) \rangle + \gamma_t d(\lambda) + B_t. \end{array}$$

Therefore,

$$\begin{array}{rcl} f_0(x_t) - \phi(\lambda_t) & \leq & \frac{1}{A_t} B_t + \min\limits_{\lambda \geq 0_m} \left\{ -\langle \lambda, f(x_t) \rangle + \frac{1}{A_t} \gamma_t d(\lambda) \right\} \\ & = & \frac{1}{A_t} B_t - \frac{A_t}{2\gamma_t} \| \left(f(x_t) \right)_+ \|_2^2. \end{array}$$

Thus,

$$f_0(x_t) + \frac{A_t}{2\gamma_t} \| \left(f(x_t) \right)_+ \|_2^2 - \phi(\lambda_t) \quad \leq \quad \frac{1}{A_t} B_t, \tag{3.7}$$

and for convergence of method (2.12) as applied to problem (3.5) we need to assume boundedness of the gradient (3.4).

3.3 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 $\mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1, \dots, n$. 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 certain acceptable level, a reasonable social target consists in arranging the production activity in accordance to the optimal solution of the following optimization problem

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^{n} \phi_i(u_i) : \sum_{i=1}^{n} P_i u_i \leq b, \; u_i \in \mathcal{U}_i, \; i = 1, \dots, n \right\}. \tag{3.8}$$

In this problem, $b \in \mathbb{R}^m_+$ is the vector of upper limits on different kind 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 (3.8) 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

$$f_i(p) \quad = \quad \max_{u_i} \left[\phi_i(u_i) - \langle p, P_i u_i \rangle : \; u_i \in \mathcal{U}_i \right], \quad i = 1, \dots, n. \tag{3.9}$$

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

$$\min_{p \geq 0} \left\{ f(p) \overset{\text{def}}{=} \langle b, p\rangle + \sum_{i=1}^{n} f_i(p) \right\} \tag{3.10}$$

This problem is dual to the optimal distribution problem (3.8). The gradient of the objective function in (3.10) is then

$$\nabla f(p) \;\; = \;\; b - v(p), \quad v(p) \overset{\text{def}}{=} \sum_{i=1}^{n} P_i u_i(p). \tag{3.11}$$

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

$$\langle \nabla f(p_*), p - p_* \rangle \;\; \geq 0, \quad \forall p \in \mathbb{R}^m_+ \tag{3.12}$$

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

The main difficulty of coordination center with solving problem (3.10) 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 (3.10) can be solved by the subgradient method with double simple averaging.

Let us present interpretation of the objects generated by method (2.18) for problem (3.10). 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:

$$s[t] \;\; = \;\; \tfrac{1}{t+1} \sum_{k=0}^{t} \nabla f(p[k]) \overset{(3.11)}{=} b - \tfrac{1}{t+1} \sum_{k=0}^{t} v(p[k]).$$

In order to apply subgradient method (2.18), we need to choose a prox-function for $\mathbb{R}^m_+$. Let us consider

$$d(p) \;\; = \;\; \tfrac{1}{2} \sum_{j=1}^{m} \tfrac{1}{\varkappa_j} (p^{(j)})^2, \tag{3.13}$$

where $\varkappa_j > 0$ are some scaling coefficients. Define $S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t} (v(p[k]) - b)$ 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 $v(p[k])$.

2. Update the aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.          (3.14)

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

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

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 $u_i[t] = \frac{1}{t+1} \sum\limits_{k=0}^{t} u_i(p[k])$, $i = 1, \dots, n$, the *historical averages* of production plans of the producers, reacting on the dynamic tax policy (3.14). Let us show that they approach the optimal solution of the socially balanced coordination problem (3.8).

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

$$
\begin{array}{rcl}
f(p[k]) + \langle \nabla f(p[k]), p - p[k] \rangle & = & \langle b, p[k] \rangle + \sum\limits_{i=1}^{n} [\phi_i(u_i(p[k])) - \langle p[k], P_i u_i(p_k) \rangle] \\[2ex]
& & + \langle b - \sum\limits_{i=1}^{n} P_i u_i(p[k]), p - p[k] \rangle \\[2ex]
& = & \sum\limits_{i=1}^{n} \phi_i(u_i(p[k])) + \langle b - v[k], p \rangle.
\end{array}
$$

Therefore,

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

Thus, in view of inequality (2.7), we have

$$
f(p[t]) - \sum\limits_{i=1}^{n} \phi_i(u_i[t]) + \frac{t+1}{\gamma_t} \sum\limits_{j=1}^{m} \frac{\varkappa_j}{2} \left(v^{(j)}[t] - b^{(j)} \right)_+^2 \quad \leq \quad \frac{1}{t+1} B_t. \tag{3.15}
$$

The left-hand side of this inequality is composed by the objective function of the dual problem (3.10), computed at the last variant of taxes $p[t]$, objective function of the primal problem (3.8), computed at historical averages $\{u_i[t]\}_{i=1}^n$, and the quadratic penalty for violation the linear inequality constraints by the historical averages:

$$v[t] - b \;\; = \;\; \sum_{i=1}^n P_i u_i[t] - b.$$

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 (3.15) will go to zero. Therefore, we come to the following conclusion.

Theorem 4 *Let taxation algorithm (3.14) apply $\gamma_t = O(\sqrt{t})$. Then the taxes $p[t]$ converge to the optimal solution of problem (3.10). At the same time, historical averages of individual production volumes $u_i[t]$, $i = 1 \dots n$, converge to the socially optimal solution of problem (3.8).*

Of course, this conclusion is valid under 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.

4 Numerical experiments

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

$$f(x) \;\; = \;\; \max\left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}. \tag{4.1}$$

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:

$$\kappa_\infty(f) \;\; \ge \;\; 2^{n+1} - 1. \tag{4.2}$$

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 $R \overset{\mathrm{def}}{=} \|x_0 - x_*\|_2 = \sqrt{n}$ and

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

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

In our experiments, we compare the simplest primal gradient method (1.2):

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

the method of simple dual averaging (1.12):

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

and the method of simple double averaging (SA$_2$; see (2.18)) with $\gamma_t = \frac{L}{R}\sqrt{t+1}$ and Euclidean prox-function $d(x) = \frac{1}{2}\|x - x_0\|_2^2$.

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). 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.

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

Table 1. Computational results for function (4.1).

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

References

[1] A. Beck, M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, **31**, 167-175 (2003)

[2] R.T. Rockafellar. *Convex Analisys.* Princeton University Press, Princeton, NJ (1970)

[3] A.S. Nemirovsky, D.B. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley and Sons, NY (1983)

[4] Yu. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer, Boston, 2004.

[5] Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120**(1), 261-283 (2009)

[6] Yu.Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, **140**(1), 125-161 (2013).

[7] B.T. Polyak. Introduction to Optimization. Software Inc., NY (1987).

[8] N.Z. Shor. *Minimization Methods for Non-differentiable Functions.* Springer-Verlag (1985).

Recent titles

CORE Discussion Papers

2013/42 Mathieu LEFEBVRE, Pierre PESTIEAU and Gregory PONTHIERE. FGT poverty measures and the mortality paradox: Theory and evidence.

2013/43 Nada BELHADJ, Jean J. GABSZEWICZ and Ornella TAROLA. Social awareness and duopoly competition.

2013/44 Volker BRITZ, P. Jean-Jacques HERINGS and Arkadi PREDTETCHINSKI. On the convergence to the Nash bargaining solution for action-dependent bargaining protocols.

2013/45 Pasquale AVELLA, Maurizio BOCCIA and Laurence WOLSEY. Single item reformulations for a vendor managed inventory routing problem: computational experience with benchmark instances.

2013/46 Alejandro LAMAS, Tanja MLINAR, Liang LU and Philippe CHEVALIER. Revenue management for operations with urgent orders.

2013/47 Helmuth CREMER, Firouz GAHVARI and Pierre PESTIEAU. Uncertain altruism and the provision of long term care.

2013/48 Claire DUJARDIN, Vincent LORANT and Isabelle THOMAS. Self-assessed health of elderly people in Brussels: does the built environment matter?

2013/49 Marc FLEURBAEY, Marie-Louise LEROUX, Pierre PESTIEAU and Grégory PONTHIERE. Fair retirement under risky lifetime.

2013/50 Manuel FÖRSTER, Ana MAULEON and Vincent VANNETELBOSCH. Trust and manipulation in social networks.

2013/51 Anthony PAPAVASILIOU, Yi HE and Alva SVOBODA. Self-commitment of combined cycle units under electricity price uncertainty.

2013/52 Ana MAULEON, Elena MOLIS, Vincent VANNETELBOSCH and Wouter VERGOTE. Dominance invariant one-to-one matching problems.

2013/53 Jean GABSZEWICZ and Skerdilajda ZANAJ. (Un)stable vertical collusive agreements.

2013/54 François MANIQUET and Massimo MORELLI. Approval quorums dominate participation quorums.

2013/55 Mélanie LEFÈVRE and Joe THARAKAN. Intermediaries, transport costs and interlinked transactions.

2013/56 Gautier M. KRINGS, Jean-François CARPANTIER and Jean-Charles DELVENNE. Trade integration and the trade imbalances in the European Union: a network perspective.

2013/57 Philip USHCHEV, Igor SLOEV and Jacques-François THISSE. Do we go shopping downtown or in the 'burbs'? Why not both?

2013/58 Mathieu PARENTI. Large and small firms in a global market: David vs. Goliath.

2013/59 Paul BELLEFLAMME and Francis BLOCH. Dynamic protection of innovations through patents and trade secrets.

2013/60 Christian HAEDO and Michel MOUCHART. Specialized agglomerations with areal data: model and detection.

2013/61 Julien MARTIN and Florian MAYNERIS. High-end variety exporters defying distance: micro facts and macroeconomic implications.

2013/62 Luca G. DEIDDA and Dimitri PAOLINI. Wage premia, education race, and supply of educated workers.

2013/63 Laurence A. WOLSEY and Hande YAMAN. Continuous knapsack sets with divisible capacities.

2013/64 Francesco DI COMITE, Jacques-François THISSE and Hylke VANDENBUSSCHE. Verti-zontal differentiation in export markets.

2013/65 Carl GAIGNÉ, Stéphane RIOU and Jacques-François THISSE. How to make the metropolitan area work? Neither big government, nor laissez-faire.

2013/66 Yu. NESTEROV and Vladimir SHIKHMAN. Algorithmic models of market equilibrium.

2013/67 Cristina PARDO-GARCIA and Jose J. SEMPERE-MONERRIS. Equilibrium mergers in a composite good industry with efficiencies.

Recent titles

CORE Discussion Papers - continued

2013/68 Federica RUSSO, Michel MOUCHART and Guillaume WUNSCH. Confounding and control in a multivariate system. An issue in causal attribution.

2013/69 Marco DI SUMMA. The convex hull of the all-different system with the inclusion property: a simple proof.

2013/70 Philippe DE DONDER and Pierre PESTIEAU. Lobbying, family concerns and the lack of political support for estate taxation.

2013/71 Alexander OSHARIN, Jacques-François THISSE, Philip USHCHEV and Valery VERBUS. Monopolistic competition and income dispersion.

2013/72 N. Baris VARDAR. Imperfect resource substitution and optimal transition to clean technologies.

2013/73 Alejandro LAMAS and Philippe CHEVALIER. Jumping the hurdles for collaboration: fairness in operations pooling in the absence of transfer payments.

2013/74 Mehdi MADANI and Mathieu VAN VYVE. A new formulation of the European day-ahead electricity market problem and its algorithmic consequences.

2014/1 Erik SCHOKKAERT and Tom TRUYTS. Preferences for redistribution and social structure.

2014/2 Maarten VAN DIJCK and Tom TRUYTS. The agricultural invasion and the political economy of agricultural trade policy in Belgium, 1875-1900.

2014/3 Ana MAULEON, Nils ROEHL and Vincent VANNETELBOSCH. Constitutions and social networks.

2014/4 Nicolas CARAYOL, Rémy DELILLE and Vincent VANNETELBOSCH. Allocating value among farsighted players in network formation.

2014/5 Yu. NESTEROV and Vladimir SHIKHMAN. Convergent subgradient methods for nonsmooth convex minimization.

Books

V. GINSBURGH and S. WEBER (2011), *How many languages make sense? The economics of linguistic diversity*. Princeton University Press.

I. THOMAS, D. VANNESTE and X. QUERRIAU (2011), *Atlas de Belgique – Tome 4 Habitat*. Academia Press.

W. GAERTNER and E. SCHOKKAERT (2012), *Empirical social choice*. Cambridge University Press.

L. BAUWENS, Ch. HAFNER and S. LAURENT (2012), *Handbook of volatility models and their applications*. Wiley.

J-C. PRAGER and J. THISSE (2012), *Economic geography and the unequal development of regions*. Routledge.

M. FLEURBAEY and F. MANIQUET (2012), *Equality of opportunity: the economics of responsibility*. World Scientific.

J. HINDRIKS (2012), *Gestion publique*. De Boeck.

M. FUJITA and J.F. THISSE (2013), *Economics of agglomeration: cities, industrial location, and globalization.* (2nd edition). Cambridge University Press.

J. HINDRIKS and G.D. MYLES (2013). *Intermediate public economics.* (2nd edition). MIT Press.

J. HINDRIKS, G.D. MYLES and N. HASHIMZADE (2013). *Solutions manual to accompany intermediate public economics.* (2nd edition). MIT Press.

CORE Lecture Series

R. AMIR (2002), Supermodularity and complementarity in economics.

R. WEISMANTEL (2006), Lectures on mixed nonlinear programming.

A. SHAPIRO (2010), Stochastic programming: modeling and theory.