CORE DISCUSSION PAPER
2005/79

Minimizing functions with bounded variation of subgradients

Yu. Nesterov *

October 2005

Abstract

In many applications it is possible to justify a reasonable bound for possible *variation* of subgradients of objective function rather than for their uniform *magnitude*. In this paper we develop a new class of efficient primal-dual subgradient schemes for such problem classes.

Keywords: convex optimization, subgradient methods, non-smooth optimization, black-box methods, lower complexity bounds.

*Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium; e-mail: nesterov@core.ucl.ac.be.

The research results presented in this 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". The scientific responsibility rests with its author.

1 Introduction

Motivation. In this paper we consider numerical schemes for solving the following problem:

$$\min_x \{f(x) : \ x \in Q\}, \tag{1.1}$$

where Q is a closed convex set in a finite dimensional space E and f is a nonsmooth convex function with $\mathrm{dom}\, f = Q$. The majority of minimization methods for solving (1.1) are justified either by imposing some bounds on the derivatives of function f:

$$\|g_x\|_* \leq L, \quad \forall g_x \in \partial f(x), \quad \forall x \in Q, \tag{1.2}$$

or, by bounding the variation of f over Q (e.g., [1], [2], [3], [6]). Note that both these measures enter as the essential factors in the worst-case complexity bounds of corresponding schemes. And both of them change if we add to the objective function a linear one.

Recently it became clear that in certain situations it is possible to get rid from such an unpleasant sensitivity. Firstly, the complexity estimates of smoothing technique [4] do not change after adding a linear function to the objective. Secondly, the dual extrapolation method [5], developed for solving the variational inequalities, can be applied to operators with *bounded variations*. Being specified for the problem (1.1), this condition looks as follows:

$$\begin{gathered} \|g_x - g_y\|_* \leq M, \\ \forall g_x \in \partial f(x), \ \forall g_y \in \partial f(y), \ \forall x, y \in Q. \end{gathered} \tag{1.3}$$

However, in both examples the approaches in use are quite special. Thus, it was not clear if it is possible to treat in a proper way the problems from the functional class (1.1), (1.3) using the standard minimization methods.

In this paper we give a positive answer on the above question. Namely, we show that the minimization schemes of Dual Averaging [7], developed for the standard setting (1.1), (1.2), can be *directly* applied to the problems satisfying a finer assumption (1.3). However, in this case we can guarantee an appropriate rate of convergence only for the primal solutions. In order to get convergence in the dual space, we need to introduce in the scheme a simple modification.

The paper is organized as follows. In Section 2 we introduce the gap functions and recall the reader the main results on dual averaging [5], [7]. In the next section we show that the simplest dual-averaging scheme as applied to the minimization problems with bounded variations of subgradients ensures a proper rate of convergence of primal variables. In Section 4 we present a simple modification of the scheme, which guarantees the right convergence for primal-dual variables.

Notations and generalities. Let E be a finite-dimensional real vector space and E^* be its dual. We denote the value of linear function $s \in E^*$ at $x \in E$ by $\langle s, x \rangle$. For measuring distances in E, let us fix some (primal) norm $\|\cdot\|$. This norm defines a system of primal balls:

$$B_r(x) = \{y \in E : \ \|y - x\| \leq r\}.$$

The dual norm $\|\cdot\|_*$ on E^* is introduced, as usual, by

$$\|s\|_* = \max_x \{\langle s, x \rangle : \ x \in B_1(0)\}, \quad s \in E^*.$$

Let Q be a closed convex set in E. Assume that we know a *prox-function* $d(x)$ of the set Q. This means that $d(x)$ is a continuous function with domain belonging to Q, which is strongly convex on Q with respect to $\|\cdot\|$: $\quad \forall x, y \in Q, \quad \forall \alpha \in [0,1]$,

$$d(\alpha x + (1-\alpha)y) \;\leq\; \alpha d(x) + (1-\alpha)d(y) - \tfrac{1}{2}\sigma\alpha(1-\alpha)\|x-y\|^2, \tag{1.4}$$

where $\sigma \geq 0$ is the *convexity parameter*. Denote by x_0 the *prox-center* of the set Q:

$$x_0 = \arg\min_x \{d(x) : \; x \in Q\}. \tag{1.5}$$

Without loss of generality, we assume that $d(x_0) = 0$. Since d is strongly convex, the prox-center is well-defined and

$$d(x) \geq \tfrac{1}{2}\sigma\|x - x_0\|^2, \quad x \in Q. \tag{1.6}$$

2 Gap functions and dual averaging

In this section we recall the reader some standard facts from [5] and [7]. We mainly follow Section 2 of the latter paper.

Let Q be a closed convex set in E endowed with a prox-function $d(x)$. We allow Q to be unbounded (for example, $Q \equiv E$). For our analysis we need to define two support-type functions of the set Q:

$$\begin{aligned} \xi_D(s) &= \max_{x \in Q}\{\langle s, x - x_0\rangle : \; d(x) \leq D\}, \\ V_\beta(s) &= \max_{x \in Q}\{\langle s, x - x_0\rangle - \beta d(x)\}, \end{aligned} \tag{2.1}$$

where $D \geq 0$ and $\beta > 0$ are some parameters. The first function is a usual support function for the set

$$\mathcal{F}_D = \{x \in Q : \; d(x) \leq D\}.$$

The second one is a proximal-type approximation of the support function of set Q. Since $d(\cdot)$ is strongly convex, for any positive D and β we have $\operatorname{dom}\xi_D = \operatorname{dom} V_\beta = E^*$. We assume Q to be simple, which means that both functions in (2.1) are computable together with their differential characteristics.

Let us mention some properties of these functions. If $\beta_2 \geq \beta_1 > 0$, then for any $s \in E^*$ we have

$$V_{\beta_2}(s) \leq V_{\beta_1}(s). \tag{2.2}$$

Note that the level of smoothness of function $V_\beta(\cdot)$ is controlled by parameter β.

Lemma 1 *Function $V_\beta(\cdot)$ is convex and differentiable on E^*. Moreover, its gradient is Lipschitz continuous with constant $\frac{1}{\beta\sigma}$:*

$$\|\nabla V_\beta(s_1) - \nabla V_\beta(s_2)\| \leq \tfrac{1}{\beta\sigma}\|s_1 - s_2\|_*, \quad \forall s_1, s_2 \in E^*. \tag{2.3}$$

For any $s \in E_*$, vector $\nabla V_\beta(s)$ belongs to Q:

$$\nabla V_\beta(s) \;=\; \pi_\beta(s) - x_0, \quad \pi_\beta(s) \;\stackrel{\text{def}}{=}\; \arg\min_{x \in Q}\{-\langle s, x\rangle + \beta d(x)\}. \tag{2.4}$$

It is important that function V is homogeneous:

$$\tau V_\beta(s) = V_{\tau\beta}(\tau s), \quad \pi_\beta(s) = \pi_{\tau\beta}(\tau s), \quad \tau > 0. \tag{2.5}$$

Lemma 2 *For any $s \in E^*$ and $\beta \geq 0$ we have*

$$\xi_D(s) \;\leq\; \beta D + V_\beta(s). \tag{2.6}$$

Consider now the sequences

$$X_k = \{x_i\}_{i=0}^k \subset Q, \quad G_k = \{g_i\}_{i=0}^k \subset E^*, \quad \Lambda_k = \{\lambda_i\}_{i=0}^k \subset R_+.$$

Typically, the test points x_i and the weights λ_i are generated by some algorithmic scheme and the points g_i are computed by a black-box oracle $\mathcal{G}(\cdot)$, related to a specific convex problem:

$$g_i = \mathcal{G}(x_i), \quad i \geq 0.$$

In this paper we consider only the problem instances, for which there exists a solution $x^* \in Q$ satisfying the condition

$$\langle g_i, x_i - x^* \rangle \geq 0, \quad i \geq 0. \tag{2.7}$$

We are going to approximate the primal and dual solutions of our problem using the following aggregate objects:

$$\begin{array}{rclcrcl} S_k &=& \sum\limits_{i=0}^k \lambda_i, & & \hat{x}_{k+1} &=& \frac{1}{S_k} \sum\limits_{i=0}^k \lambda_i x_i, \\ s_{k+1} &=& \sum\limits_{i=0}^k \lambda_i g_i, & & \hat{s}_{k+1} &=& \frac{1}{S_k} s_{k+1}, \end{array} \tag{2.8}$$

with $\hat{x}_0 = x_0$ and $s_0 = 0$.

As we will see later, the quality of the test sequence X_k can be naturally described by the following *gap function*:

$$\delta_k(D) = \max_x \left\{ \sum_{i=0}^k \lambda_i \langle g_i, x_i - x \rangle : \; x \in \mathcal{F}_D, \; \right\}, \quad D \geq 0. \tag{2.9}$$

Using notation (2.8), we get an explicit representation of the gap:

$$\delta_k(D) = \sum_{i=0}^k \lambda_i \langle g_i, x_i - x_0 \rangle + \xi_D(-s_{k+1}). \tag{2.10}$$

Sometimes we will use an *upper gap function*

$$\begin{array}{rcl} \Delta_k(\beta, D) &=& \beta D + \sum\limits_{i=0}^k \lambda_i \langle g_i, x_i - x_0 \rangle + V_\beta(-s_{k+1}) \\ &=& \sum\limits_{i=0}^k \lambda_i \langle g_i, x_i - \pi_\beta(-s_{k+1}) \rangle + \beta \cdot (D - d(\pi_\beta(-s_{k+1}))) \, . \end{array} \tag{2.11}$$

In view of (2.6) and (2.10), for any non-negative D and β we have

$$\delta_k(D) \leq \Delta_k(\beta, D). \tag{2.12}$$

Since Q is a simple set, the values of the gap functions can be easily computed. Note that for some D these values can be negative. However, if the solution x^* of our problem do exist (in the sense of (2.7)), then for

$$D \geq d(x^*)$$

the value $\delta_k(D)$ is non-negative independently on the sequences X_k, Λ_k and G_k, involved in its definition.

Consider now the generic scheme of *Dual Averaging* (DA-scheme) [7].

Initialization: Set $s_0 = 0 \in E^*$. Choose $\beta_0 > 0$.

Iteration ($k \geq 0$):

1. Compute $g_k = \mathcal{G}(x_k)$.

2. Choose $\lambda_k > 0$ and set $s_{k+1} = s_k + \lambda_k g_k$.

3. Choose $\beta_{k+1} > 0$ and set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$. (2.13)

Theorem 1 *Let the sequences X_k, G_k and Λ_k be generated by (2.13), and the parameters $\{\beta_k\}_{k=0}^{\infty}$ satisfy the condition*

$$\beta_{k+1} \geq \beta_k, \quad k \geq 0. \tag{2.14}$$

Then:

1. *For any $k \geq 0$ and $D \geq 0$ we have:*

$$\delta_k(D) \;\leq\; \Delta_k(\beta_{k+1}, D) \;\leq\; \beta_{k+1} D + \tfrac{1}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.15}$$

2. *Assume that the solution x^* in the sense (2.7) exists. Then*

$$\tfrac{1}{2}\sigma \|x_{k+1} - x^*\|^2 \leq d(x^*) + \tfrac{1}{2\sigma\beta_{k+1}} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.16}$$

Thus, if the scheme (2.13) is applied to the problem (1.1) with oracle $\mathcal{G}(x) \in \partial f(x)$, which satisfies condition (1.2), then inequality (2.15) establishes the rate of convergence of the generated sequence. Indeed, if $D \geq d(x^*)$, then in view of definition of the gap function (2.9), we have

$$\begin{aligned} \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i f(x_i) - f(x^*) \;&\leq\; \tfrac{1}{S_k}\delta_k(D) \;\leq\; \tfrac{1}{S_k}\Delta_k(\beta_{k+1}, D) \\ &\leq\; \tfrac{1}{S_k}\left[\beta_{k+1} D + \tfrac{L^2}{2\sigma} \sum_{i=0}^{k} \tfrac{\lambda_i^2}{\beta_i}\right]. \end{aligned} \tag{2.17}$$

In the scheme of *Simple Averages* (SA) we take

$$\lambda_k \equiv 1, \quad \beta_0 = \gamma, \quad \beta_k = \gamma\sqrt{k}, \quad k \geq 1, \tag{2.18}$$

where γ is a positive parameter. Then, from (2.17) we derive the following bound:

$$\begin{aligned} \frac{1}{k+1}\sum_{i=0}^{k} f(x_i) - f(x^*) &\leq \frac{1}{k+1}\left[\gamma D\sqrt{k+1} + \frac{L^2}{2\gamma\sigma}\left(1 + \sum_{i=1}^{k}\frac{1}{\sqrt{i}}\right)\right] \\ &\leq \frac{1}{\sqrt{k+1}}\left[\gamma D + \frac{L^2}{\gamma\sigma}\right]. \end{aligned} \tag{2.19}$$

3 Bounded variation of subgradients

Let us show now how to bound the gap functions using the assumption (1.3). Denote $\mu_k \stackrel{\text{def}}{=} \frac{\beta_{k+1}}{S_k}$. Then

$$\begin{aligned} \Psi_k &\stackrel{\text{def}}{=} \frac{1}{S_k}\Delta_k(\beta_{k+1}, D) \\ &\stackrel{(2.11)}{=} \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i\langle g_i, x_i - x_0\rangle + \mu_k D + \frac{1}{S_k}V_{\beta_{k+1}}(-s_{k+1}) \\ &\stackrel{(2.5),(2.8)}{=} \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i\langle g_i, x_i - x_0\rangle + \mu_k D + V_{\mu_k}(-\hat{s}_{k+1}). \end{aligned} \tag{3.1}$$

Let us introduce for scheme (2.13) two new requirements which are assumed to be valid for all $k \geq 0$:

$$\begin{aligned} &\text{A1.} \quad \mu_{k+1} \leq \mu_k. \\ &\text{A2.} \quad d(x_k) \leq D. \end{aligned} \tag{3.2}$$

Note that the first requirement is satisfied by the strategy (2.18). The second assumption can be valid, for example, due to the boundedness of the set Q. Or, if we ensure condition (2.14), it can be derived from inequality (2.16).

In view of (3.1), we have

$$\begin{aligned} \Psi_{k+1} &\stackrel{(2.11)}{=} \frac{1}{S_{k+1}}\sum_{i=0}^{k+1}\lambda_i\langle g_i, x_i - x_{k+2}\rangle + \mu_{k+1}\cdot(D - d(x_{k+2})) \\ &\stackrel{(3.2)}{\leq} \frac{1}{S_{k+1}}\sum_{i=0}^{k+1}\lambda_i\langle g_i, x_i - x_{k+2}\rangle + \mu_k\cdot(D - d(x_{k+2})) \\ &\stackrel{(2.8)}{=} \frac{1}{S_{k+1}}\sum_{i=0}^{k+1}\lambda_i\langle g_i, x_i - x_0\rangle - \langle\hat{s}_{k+2}, x_{k+2} - x_0\rangle + \mu_k\cdot(D - d(x_{k+2})) \\ &\stackrel{(2.1)}{\leq} \frac{1}{S_{k+1}}\sum_{i=0}^{k+1}\lambda_i\langle g_i, x_i - x_0\rangle + \mu_k D + V_{\mu_k}(-\hat{s}_{k+2}). \end{aligned} \tag{3.3}$$

Denote $\alpha_k = \frac{\lambda_{k+1}}{S_{k+1}}$. Then

$$\begin{aligned}
\tfrac{1}{S_{k+1}} \sum_{i=0}^{k+1} \lambda_i \langle g_i, x_i - x_0 \rangle &= (1-\alpha_k) \tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0 \rangle + \alpha_k \langle g_{k+1}, x_{k+1} - x_0 \rangle, \\
\tfrac{1}{S_k} \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0 \rangle &\overset{(3.1)}{=} \Psi_k - \mu_k D - V_{\mu_k}(-\hat{s}_{k+1}), \\
\langle g_{k+1}, x_{k+1} - x_0 \rangle &\overset{(2.13)}{=} \langle g_{k+1}, \pi_{\beta_{k+1}}(-s_{k+1})) - x_0 \rangle \\
&\overset{(2.5)}{=} \langle g_{k+1}, \pi_{\mu_k}(-\hat{s}_{k+1})) - x_0 \rangle \\
&= \langle g_{k+1}, \nabla V_{\mu_k}(-\hat{s}_{k+1}) \rangle.
\end{aligned}$$

Substituting these equalities in (3.3), we obtain:

$$\begin{aligned}
\Psi_{k+1} &\leq \mu_k D + V_{\mu_k}(-\hat{s}_{k+2}) + (1-\alpha_k)\left[\Psi_k - \mu_k D - V_{\mu_k}(-\hat{s}_{k+1})\right] \\
&\quad + \alpha_k \langle g_{k+1}, \nabla V_{\mu_k}(-\hat{s}_{k+1}) \rangle \\
&= (1-\alpha_k)\Psi_k + \alpha_k \mu_k D \\
&\quad + \left[V_{\mu_k}(-\hat{s}_{k+2}) - (1-\alpha_k) V_{\mu_k}(-\hat{s}_{k+1}) + \alpha_k \langle g_{k+1}, \nabla V_{\mu_k}(-\hat{s}_{k+1}) \rangle\right].
\end{aligned} \tag{3.4}$$

Note that

$$\begin{aligned}
\hat{s}_{k+2} &= \tfrac{1}{S_k + \lambda_{k+1}} \left[S_k \cdot \hat{s}_{k+1} + \lambda_{k+1} g_{k+1}\right] \\
&= (1-\alpha_k)\hat{s}_{k+1} + \alpha_k g_{k+1} \\
&= \hat{s}_{k+1} + \alpha_k (g_{k+1} - \hat{s}_{k+1}).
\end{aligned}$$

Hence, in view of (2.3), we can estimate

$$V_{\mu_k}(-\hat{s}_{k+2}) \leq V_{\mu_k}(-\hat{s}_{k+1}) - \alpha_k \langle g_{k+1} - \hat{s}_{k+1}, \nabla V_{\mu_k}(-\hat{s}_{k+1}) \rangle + \tfrac{\alpha_k^2}{2\sigma\mu_k} \|g_{k+1} - \hat{s}_{k+1}\|_*^2.$$

Hence, for the expression in brackets in (3.4) we get an upper bound

$$[\,\cdot\,] \leq \alpha_k V_{\mu_k}(-\hat{s}_{k+1}) + \alpha_k \langle \hat{s}_{k+1}, \nabla V_{\mu_k}(-\hat{s}_{k+1}) \rangle + \tfrac{\alpha_k^2}{2\sigma\mu_k} \|g_{k+1} - \hat{s}_{k+1}\|_*^2.$$

Since $V_\mu(\cdot)$ is a convex function and $V_\mu(0) = 0$, the latter inequality together with (3.4) result in the following statement.

Lemma 3 *Assume that the objects involved in the scheme of Dual Averaging (2.13) satisfy conditions (3.2). Then for any $k \geq 0$ we have*

$$\Psi_{k+1} \leq (1-\alpha_k)\Psi_k + \alpha_k \mu_k D + \tfrac{\alpha_k^2}{2\sigma\mu_k} \|g_{k+1} - \hat{s}_{k+1}\|_*^2. \tag{3.5}$$

Now we can estimate the rate of convergence of SA-method (2.13), (2.18) as applied to the problem (1.1) with oracle $\mathcal{G}(x) \in \partial f(x)$.

Theorem 2 *Let problem (1.1) satisfy assumption (1.3), and sequence $\{x_k\}_{k=0}^N$ generated by SA-method (2.13), (2.18) satisfy condition A2 in (3.2) with $D \geq d(x^*)$. Then for any $N \geq 1$ we have*

$$\frac{1}{N} \sum_{k=1}^{N} f(x_k) - f^* \quad \leq \quad \frac{2}{\sqrt{N}} \left[\gamma D + \frac{1}{2\gamma\sigma} M^2 \right]. \tag{3.6}$$

Proof:
Note that for $k \geq 0$ the choice (2.18) results in

$$S_k = k+1, \quad \alpha_k = \frac{\lambda_{k+1}}{S_{k+1}} = \frac{1}{k+2}, \quad \mu_k = \frac{\beta_{k+1}}{S_k} = \frac{\gamma}{\sqrt{k+1}}. \tag{3.7}$$

Hence, condition A1 in (3.2) is valid.

Further, by definition (2.8) we have $\hat{s}_{k+1} \in \mathrm{Conv}\,\{g_0, \ldots, g_k\}$. Therefore, in view of assumption (1.3), we can bound

$$\|g_{k+1} - \hat{s}_{k+1}\|_* \leq M, \quad k \geq 0.$$

Applying now the statement of Lemma 3, we obtain

$$\begin{aligned} \Psi_{k+1} \quad &\leq \quad \frac{k+1}{k+2} \Psi_k + \frac{\gamma D}{(k+2)\sqrt{k+1}} + \frac{M^2 \sqrt{k+1}}{2\gamma\sigma(k+2)^2} \\ &\leq \quad \frac{k+1}{k+2} \Psi_k + \frac{1}{(k+2)\sqrt{k+1}} \left[\gamma D + \frac{1}{2\gamma\sigma} M^2 \right]. \end{aligned}$$

Denoting now $\tau_k = (k+1)\Psi_k$ and $C = \gamma D + \frac{1}{2\gamma\sigma} M^2$, we come to recursive bounds

$$\tau_{k+1} \leq \tau_k + \frac{1}{\sqrt{k+1}} C, \quad k \geq 0.$$

Hence, for any $N \geq 1$

$$\tau_N \leq \tau_0 + C \sum_{k=1}^{N} \frac{1}{\sqrt{k}} \leq \tau_0 + (2\sqrt{N} - 1) C. \tag{3.8}$$

Further, for any $x \in Q$ denote by $f'(x)$ an arbitrary element from $\partial f(x)$. By definition (3.1), we have

$$\begin{aligned} \tau_0 \quad &= \quad \Psi_0 \;=\; \mu_0 D + V_{\mu_0}(-\hat{s}_1) \quad \overset{(2.8),(3.7)}{=} \quad \gamma D + V_\gamma(-g_0) \\ &\overset{(2.1)}{=} \quad \gamma D + \max_{x \in Q} \{ \langle g_0, x_0 - x \rangle - \gamma d(x) \} \\ &= \quad \gamma D + \max_{x \in Q} \{ \langle f'(x), x_0 - x \rangle + \langle g_0 - f'(x), x_0 - x \rangle - \gamma d(x) \} \qquad (3.9) \\ &\overset{(1.3),(1.6)}{\leq} \quad \gamma D + \max_{x \in Q} \{ \langle f'(x), x_0 - x \rangle + M \|x - x_0\| - \tfrac{1}{2} \gamma\sigma \|x - x_0\|^2 \} \\ &\leq \quad \gamma D + \tfrac{1}{2\gamma\sigma} M^2 + f(x_0) - f(x^*). \end{aligned}$$

Thus, substituting this estimate in (3.8), we obtain

$$\begin{aligned} f(x_0) - f(x^*) + 2\sqrt{N}C \quad &\geq \quad \tau_N \;=\; (N+1)\Psi_N \overset{(3.1)}{=} \Delta_N(\beta_{N+1}, D) \\ &\overset{(2.12)}{\geq} \quad \delta_N(D) \overset{(2.9)}{=} \max_x \left\{ \sum_{k=0}^{N} \langle g_k, x_k - x \rangle : \; x \in \mathcal{F}_D \right\} \\ &\geq \quad \max_x \left\{ \sum_{k=0}^{N} [f(x_k) - f(x)] : \; x \in \mathcal{F}_D \right\} \\ &= \quad \sum_{k=0}^{N} f(x_k) - (N+1) f(x^*). \end{aligned}$$

Clearly, this inequality can be rewritten in the form (3.6). □

It is interesting that the value $f(x_0)$ cannot appear in the left-hand side of inequality (3.6). Indeed, an augmentation of the objective function by a linear term can change the gap $f(x_0) - f(x^*)$ in an arbitrary way. On the other hand, such a modification does not affect anyhow the right-hand side of inequality (3.6). Thus, we can see that the influence of a "linear part" of the objective function of problem (1.1) is eliminated by SA-method in one iteration.

Note also that we managed to get a rate of convergence only for the sequence of the values of objective function. However, as it was shown in [7], the rate of convergence for approximate *primal-dual* solutions of optimization problems can be derived only from the convergence rate of the gap functions. In the proof of Theorem 2 we have seen that the upper bound for this gap (3.9) includes residual $f(x_0) - f(x^*)$, which does not admit any upper bound in terms of M. Nevertheless, in the next section we show that our goal can be achieved by a simple modification of the general DA-scheme (2.13).

4 General problems

Note that the technique presented in Section 3 can be adapted to all classes of problems considered in [7] provided that the variation of the answers of their oracles is bounded. However, as we have seen in the previous section, the rules of DA-methods need some modifications. Namely, the information related to the starting point x_0 must be treated in a very special way.

In order to describe a general scheme of modified DA-methods, we need to introduce notation for some new objects. All of them can be seen as truncated versions of

corresponding prototypes introduced in Section 2:

$$\begin{aligned} S_k^+ &= \sum_{i=1}^k \lambda_i, \quad \hat{x}_k^+ = \tfrac{1}{S_k^+} \sum_{i=1}^k \lambda_i x_i, \quad s_k^+ = \sum_{i=1}^k \lambda_i g_i, \quad \hat{s}_k^+ = \tfrac{1}{S_k^+} s_k^+, \\ \delta_k^+(D) &= \max_x \left\{ \sum_{i=1}^k \lambda_i \langle g_i, x_i - x \rangle : \; x \in \mathcal{F}_D, \; \right\}, \quad D \geq 0, \\ \Delta_k^+(\beta, D) &= \beta D + \sum_{i=1}^k \lambda_i \langle g_i, x_i - x_0 \rangle + V_\beta(-s_k^+) \\ &= \sum_{i=1}^k \lambda_i \langle g_i, x_i - \pi_\beta(-s_k^+) \rangle + \beta \cdot \Big(D - d(\pi_\beta(-s_k^+)) \Big). \end{aligned} \tag{4.1}$$

As before, for any non-negative β and D we have

$$\delta_k^+ \; \leq \; \Delta_k^+(\beta, D). \tag{4.2}$$

Consider the following scheme of *Truncated Dual Averaging* (TDA). Below we choose

$$\lambda_1 = \lambda_0, \quad \beta_1 = \beta_0. \tag{4.3}$$

Initialization: Set $s_0^+ = 0 \in E^*$.

Choose $\beta_0 > 0$. Set $x_1 = \pi_{\beta_0}(-\lambda_0 g_0)$.

Iteration ($k \geq 1$):

1. Compute $g_k = \mathcal{G}(x_k)$.

2. Choose $\lambda_k > 0$ and set $s_k^+ = s_{k-1}^+ + \lambda_k g_k$.

3. Choose $\beta_k > 0$ and set $x_{k+1} = \pi_{\beta_k}(-s_k^+)$. (4.4)

As compared with (2.13), in the modified scheme we exclude x_0 and g_0 from all aggregated objects. In what follows we denote $\mu_k^+ = \beta_k / S_k^+$.

Theorem 3 *Assume the variation of the answers of oracle $\mathcal{G}$ be bounded:*

$$\begin{aligned} \|g_x - g_y\|_* \leq M, \\ \forall g_x \in \mathcal{G}(x), \; \forall g_y \in \mathcal{G}(y), \; \forall x, y \in Q. \end{aligned} \tag{4.5}$$

Let sequences $\{x_k\}_{k=1}^\infty$, $\{g_k\}_{k=1}^\infty$ and $\{\lambda_k\}_{k=1}^\infty$ be generated by (4.4). Assume that

$$d(x_k) \leq D, \quad \mu_{k+1}^+ \leq \mu_k^+, \quad k \geq 1. \tag{4.6}$$

Then, for any $k \geq 1$ and $D \geq 0$ we have:

$$\delta_k^+(D) \;\leq\; \Delta_k^+(\beta_k, D) \;\leq\; D\left[\beta_0 + \sum_{i=2}^{k} \frac{\lambda_i \beta_{i-1}}{S_{i-1}^+}\right] + \frac{1}{2\sigma} M^2 \sum_{i=1}^{k} \frac{\lambda_i^2}{\beta_{i-1}}. \tag{4.7}$$

Proof:
Denote $\alpha_k = \lambda_{k+1}/S_{k+1}^+$. Then, for any $k \geq 1$ we have

$$\begin{aligned}
\Delta_{k+1}^+(\beta_{k+1}, D) &= \beta_{k+1} D + V_{\beta_{k+1}}(-s_{k+1}^+) + \sum_{i=1}^{k+1} \lambda_i \langle g_i, x_i - x_0 \rangle \\
&= \beta_{k+1} D + V_{\beta_{k+1}}(-s_{k+1}^+) + \lambda_{k+1} \langle g_{k+1}, x_{k+1} - x_0 \rangle \\
&\quad + \Delta_k^+(\beta_k, D) - \beta_k D - V_{\beta_k}(-s_k^+) \qquad (4.8) \\
&\overset{(2.5)}{=} S_{k+1}^+ \left[\mu_{k+1}^+ D + V_{\mu_{k+1}^+}(-\hat{s}_{k+1}^+) + \alpha_k \langle g_{k+1}, x_{k+1} - x_0 \rangle\right] \\
&\quad + \Delta_k^+(\beta_k, D) - \beta_k D - S_k^+ V_{\mu_k^+}(-\hat{s}_k^+).
\end{aligned}$$

In view of (2.5) and assumption (4.6), we have

$$\begin{aligned}
\mu_{k+1}^+ D + V_{\mu_{k+1}^+}(-\hat{s}_{k+1}^+) &= \mu_{k+1}^+ D + \langle -\hat{s}_{k+1}^+, x_{k+2} - x_0 \rangle - \mu_{k+1}^+ d(x_{k+2}) \\
&\leq \mu_k^+ D + \langle -\hat{s}_{k+1}^+, x_{k+2} - x_0 \rangle - \mu_k^+ d(x_{k+2}) \\
&\leq \mu_k^+ D + V_{\mu_k^+}(-\hat{s}_{k+1}^+).
\end{aligned}$$

Since $S_k^+ = (1 - \alpha_k) S_{k+1}^+$, and $S_{k+1}^+ \mu_k^+ = \frac{\beta_k}{1 - \alpha_k}$, from (4.8) we obtain

$$\begin{aligned}
\Delta_{k+1}^+(\beta_{k+1}, D) &\leq \Delta_k^+(\beta_k, D) + \tfrac{\alpha_k \beta_k}{1 - \alpha_k} D \\
&\quad + S_{k+1}^+ [V_{\mu_k^+}(-\hat{s}_{k+1}^+) + \alpha_k \langle g_{k+1}, x_{k+1} - x_0 \rangle - (1 - \alpha_k) V_{\mu_k^+}(-\hat{s}_k^+)].
\end{aligned}$$

Note that $\hat{s}_{k+1}^+ = (1 - \alpha_k)\hat{s}_k^+ + \alpha_k g_{k+1} \in \operatorname{Conv}\{g_1, \ldots, g_{k+1}\}$. Since

$$\nabla V_{\mu_k^+}(-\hat{s}_k^+) = x_{k+1} - x_0,$$

in view of (2.3) and assumption (4.5) we have

$$\begin{aligned}
V_{\mu_k^+}(-\hat{s}_{k+1}^+) &\leq V_{\mu_k^+}(-\hat{s}_k^+) + \alpha_k \langle \hat{s}_k^+ - g_{k+1}, \nabla V_{\mu_k^+}(-\hat{s}_k^+) \rangle + \tfrac{\alpha_k^2}{2\sigma\mu_k^+} \|\hat{s}_k^+ - g_{k+1}\|_*^2 \\
&\leq V_{\mu_k^+}(-\hat{s}_k^+) + \alpha_k \langle \hat{s}_k^+ - g_{k+1}, x_{k+1} - x_0 \rangle + \tfrac{\alpha_k^2}{2\sigma\mu_k^+} M^2.
\end{aligned}$$

Hence, we can continue:

$$\begin{aligned}
\Delta^+_{k+1}(\beta_{k+1}, D) &\leq \Delta^+_k(\beta_k, D) + \tfrac{\alpha_k\beta_k}{1-\alpha_k} D \\
&\quad + \alpha_k S^+_{k+1}\left[V_{\mu^+_k}(-\hat{s}^+_k) + \langle \hat{s}^+_k, x_{k+1} - x_0\rangle + \tfrac{\alpha_k}{2\sigma\mu^+_k} M^2\right] \\
&\leq \Delta^+_k(\beta_k, D) + \tfrac{\alpha_k\beta_k}{1-\alpha_k} D + \tfrac{\alpha_k^2 S^+_{k+1}}{2\sigma\mu^+_k} M^2 \\
&\leq \Delta^+_k(\beta_k, D) + \tfrac{\lambda_{k+1}\beta_k}{S^+_k} D + \tfrac{\lambda^2_{k+1}}{2\sigma\beta_k} M^2.
\end{aligned}$$

Thus, for any $k \geq 2$ we get the following bound:

$$\Delta^+_k(\beta_k, D) \leq \Delta^+_1(\beta_1, D) + D \sum_{i=2}^{k} \tfrac{\lambda_i\beta_{i-1}}{S^+_{i-1}} + \tfrac{1}{2\sigma} M^2 \sum_{i=2}^{k} \tfrac{\lambda_i^2}{\beta_{i-1}}. \tag{4.9}$$

It remains to estimate the first term in the right-hand side of this inequality. Note that

$$\begin{aligned}
\Delta^+_1(\beta_1, D) &= \beta_1 D + \lambda_1\langle g_1, x_1 - x_0\rangle + V_{\beta_1}(-\lambda_1 g_1) \\
&= \beta_1 D + \lambda_1\langle g_1, \nabla V_{\beta_0}(-\lambda_0 g_0)\rangle + V_{\beta_1}(-\lambda_1 g_1) \qquad (4.10) \\
&\overset{(4.3)}{=} \beta_0 D + \lambda_1\langle g_1, \nabla V_{\beta_0}(-\lambda_1 g_1)\rangle + V_{\beta_0}(-\lambda_1 g_1).
\end{aligned}$$

However, in view of (2.3), we have

$$\begin{aligned}
V_{\beta_0}(-\lambda_1 g_1) &\overset{(4.5)}{\leq} V_{\beta_0}(-\lambda_1 g_0) + \lambda_1\langle g_0 - g_1, \nabla V_{\beta_0}(-\lambda_1 g_0)\rangle + \tfrac{\lambda_1^2}{2\sigma\beta_0} M^2 \\
(V_\beta(\cdot) \text{ is convex; } V_\beta(0) = 0) &\leq -\lambda_1\langle g_1, \nabla V_{\beta_0}(-\lambda_1 g_0)\rangle + \tfrac{\lambda_1^2}{2\sigma\beta_0} M^2.
\end{aligned}$$

Thus, using this inequality in (4.10), we get

$$\Delta^+_1(\beta_1, D) \leq \beta_0 D + \tfrac{\lambda_1^2}{2\sigma\beta_0} M^2.$$

Now, taking into account (4.9), we obtain (4.7). □

Thus, using TDA-scheme (4.4) we can guarantee a certain rate of convergence of the gap function $\frac{1}{S^+_k}\Delta^+_k(\beta_k, D)$ to zero. This means that corresponding methods are able to generate the approximate dual solutions (see [7] for details). Note that the strategy (2.18) leads to the following variant of the estimate (4.7):

$$\begin{aligned}
\tfrac{1}{S^+_k}\delta^+_k(D) \leq \tfrac{1}{S^+_k}\Delta^+_k(\beta_k, D) &\leq \tfrac{1}{k} D\gamma\left[1 + \sum_{i=2}^{k} \tfrac{1}{\sqrt{i-1}}\right] + \tfrac{1}{2\sigma\gamma k} M^2\left[1 + \sum_{i=2}^{k} \tfrac{1}{\sqrt{i-1}}\right] \\
&\leq \tfrac{2}{\sqrt{k}}\left[D\gamma + \tfrac{1}{2\sigma\gamma} M^2\right], \quad k \geq 1.
\end{aligned}$$

References

[1] A. Ben-Tal, T. Margalit, A. Nemirovski. The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography. *SIAM Journal on Optimization*. **12**, 79-108, (2001).

[2] Hiriart-Urruty, J.-B., and Lemarechal, C.: Convex Analysis and Minimization Algorithms. Vols. I and II, (Springer-Verlag, Berlin and N. Y., 1993).

[3] C. Lemarechal, A. Nemirovskii and Yu. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69 (1995), 111-147.

[4] Yu.Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming (A), **103**(1), 127-152, (2005).

[5] Yu. Nesterov. Dual extrapolation and its applications for solving variational inequalities and related problems. CORE Discussion Paper #2003/68, CORE 2003. Accepted by *Mathematical Programming* (B).

[6] Yu. Nesterov. Introductory Lectures on Convex Optimization. A basic course. (Kluwer, Boston, 2004).

[7] Yu. Nesterov. Primal-dual subgradient methods for convex problems. CORE DP 2005/67, September 2005.