

# EXCESSIVE GAP TECHNIQUE IN NON-SMOOTH CONVEX MINIMIZATION\*

YU. NESTEROV<sup>†</sup>

**Abstract.** In this paper we introduce a new primal-dual technique for convergence analysis of gradient schemes for non-smooth convex optimization. As an example of its application, we derive a primal-dual gradient method for a special class of structured non-smooth optimization problems, which ensures a rate of convergence of the order  $O(\frac{1}{k})$ , where  $k$  is the iteration count. Another example is a gradient scheme, which minimizes a non-smooth strongly convex function with known structure with the rate of convergence  $O(\frac{1}{k^2})$ . In both cases the efficiency of the methods is higher than the corresponding black-box lower complexity bounds by an order of magnitude.

**Key words.** Convex optimization, non-smooth optimization, complexity theory, black-box oracle, optimal methods, structural optimization.

**AMS subject classifications.** 90C25, 90C47, 68Q25.

**1. Introduction.** This paper continues the research started in [3], where it was shown that some structured non-smooth optimization problems can be solved with efficiency estimates  $O(\frac{1}{\epsilon})$ , where  $\epsilon$  is the desired accuracy of the solution. This complexity is much better than the theoretical lower complexity bound  $O(\frac{1}{\epsilon^2})$  (see [2]). This improvement, of course, is possible because of certain relaxation of the standard black box assumption. Instead, it was assumed that our problem has an explicit and quite simple minimax structure. The numerical scheme proposed in [3] has a drawback, which decreases its practical efficiency. In this scheme the number of steps must be fixed in advance. It is chosen in accordance to a worst case complexity analysis.

In this paper we propose several new primal-dual gradient schemes for the same class of problems as in [3]. However, our schemes now are free from the above deficiency. They are derived from a new primal-dual symmetric technique for convergence analysis, which we call the *excessive gap condition*.

The paper is organized as follows. In Section 2 we introduce our *model* of optimization problem and recall several useful facts from [3]. In Section 3 we describe the excessive gap condition. In the next two sections we present two different strategies for maintaining the condition during the optimization process. In Section 6 we give the convergence result of the order  $O(\frac{1}{k})$ , where  $k$  is the iteration counter. This convergence result is valid for all non-smooth functions, described by our model. However, if we assume more, (namely, the strong convexity of the primal objective), then the convergence can be improved up to  $O(\frac{1}{k^2})$ . This improvement is presented in the last Section 7. Note that both complexity results improve the corresponding general lower complexity bound by an order of magnitude.

In what follows we work with different primal and dual spaces equipped by corresponding norms. For sake of notation, we apply the following convention. The (primal) finite-dimensional real vector space is always denoted by  $E$ , possibly with an index. This space is endowed with a norm  $\|\cdot\|$ , which has the same index as the corresponding space. The space of linear functions on  $E$  is denoted by  $E^*$ . For  $s \in E^*$  and  $x \in E$  we denote  $\langle s, x \rangle$  the value of  $s$  at  $x$ . The *scalar product*  $\langle \cdot, \cdot \rangle$  is marked by the same index as  $E$ . The norm for the dual space is defined in the standard way:

$$\|s\|^* = \max_{\|x\|=1} \langle s, x \rangle.$$

For operator  $A : E_1 \rightarrow E_2^*$  we define the *adjoint* operator  $A^* : E_2 \rightarrow E_1^*$  by identity

$$\langle Ax, u \rangle_2 \equiv \langle A^*u, x \rangle_1 \quad \forall x \in E_1, u \in E_2.$$

The *norm* of such an operator is defined as follows:

$$\|A\|_{1,2} = \max_{\|x\|_1=1} \max_{\|u\|_2=1} \langle Ax, u \rangle_2.$$

---

\* This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the author.

<sup>†</sup>Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium ([neesterov@core.ucl.ac.be](mailto:neesterov@core.ucl.ac.be)).

Clearly,

$$\|A\|_{1,2} = \|A^*\|_{2,1} = \max_{\|x\|_1=1} \|Ax\|_2^* = \max_{\|u\|_2=1} \|A^*u\|_1^*.$$

Hence, for any  $h \in E_1$  we have

$$(1.1) \quad \|Ah\|_2^* \leq \|A\|_{1,2} \cdot \|h\|_1.$$

Further, recall that function  $d(x)$  is called *strongly convex* on a closed convex set  $Q$  if for any  $\alpha \in [0, 1]$  we have

$$d(\alpha x + (1 - \alpha)y) \leq \alpha d(x) + (1 - \alpha)d(y) - \frac{1}{2}\alpha(1 - \alpha)\sigma\|x - y\|^2, \quad x, y \in Q.$$

In this inequality, the constant  $\sigma$  is called the (*strong*) *convexity parameter* of  $d$ . We often use the following property of strongly convex functions:

$$(1.2) \quad d(x) \geq d(x_*) + \frac{1}{2}\sigma\|x - x_*\|^2, \quad x \in Q,$$

where  $x_* = \arg \min_{x \in Q} d(x)$ . If  $d$  is differentiable, the equivalent definitions of strong convexity are as follows (see, for example, Section 2.1.3 [4]):

$$(1.3) \quad d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\sigma\|y - x\|_1^2, \quad x, y \in Q,$$

$$(1.4) \quad \langle \nabla d(x) - \nabla d(y), x - y \rangle \geq \sigma\|x - y\|^2, \quad x, y \in Q.$$

Finally, we say that function  $f(x)$  has Lipschitz continuous gradient on a convex set  $Q$  if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \quad x, y \in Q,$$

for some constant  $L \geq 0$ . Then

$$(1.5) \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2, \quad x, y \in Q,$$

(see, for example, Section 2.1.1 [4]).

**2. A class of structured problems.** In this paper we are interested in the following minimization problem:

$$(2.1) \quad \text{Find } f^* = \min_{x \in Q_1} f(x),$$

where  $Q_1$  is a bounded closed convex set in a finite-dimensional real vector space  $E_1$  and  $f(x)$  is a continuous convex function on  $Q_1$ . We do not assume  $f$  to be differentiable.

Very often, the *structure* of the objective function in (2.1) is known. Let us assume that this structure can be described by the following *model* (see [3] for different examples):

$$(2.2) \quad f(x) = \hat{f}(x) + \max_{u \in Q_2} \{\langle Ax, u \rangle_2 - \hat{\phi}(u)\},$$

where function  $\hat{f}(x)$  is continuous and convex on  $Q_1$ ,  $Q_2$  is a closed convex bounded set in a finite-dimensional real vector space  $E_2$ ,  $\hat{\phi}(u)$  is a continuous convex function on  $Q_2$  and the linear operator  $A$  maps  $E_1$  to  $E_2^*$ . In this case the problem (2.1) can be written in an *adjoint* form:

$$(2.3) \quad \max_{u \in Q_2} \phi(u),$$

$$\phi(u) = -\hat{\phi}(u) + \min_{x \in Q_1} \{\langle Ax, u \rangle_2 + \hat{f}(x)\}.$$

We assume that this representation is completely similar to (2.1) in the following sense. The methods described in this paper are implementable only if the optimization problems involved in the definitions of functions  $f(x)$  and  $\phi(u)$  can be solved in a closed form. So, we assume that the structure of the objects  $\hat{f}$ ,  $\hat{\phi}$ ,  $Q_1$  and  $Q_2$  is simple enough. We also assume that the functions  $\hat{f}$  and  $\hat{\phi}$  have Lipschitz-continuous gradients with Lipschitz constants  $L_1(\hat{f})$  and  $L_2(\hat{\phi})$  respectively.

Let us show that the knowledge of structure (2.2) can help in solving problems (2.1) and (2.3). As in [3], we are going to use this structure for constructing a smooth approximation of the objective functions.

Consider a *prox-function*  $d_2(u)$  of the set  $Q_2$ . This means that  $d_2(u)$  is continuous and strongly convex on  $Q_2$  with a strong convexity parameter  $\sigma_2 > 0$ . Denote by

$$u_0 = \arg \min_{u \in Q_2} d_2(u)$$

the *prox-center* of the function  $d_2(\cdot)$ . Without loss of generality we assume that  $d_2(u_0) = 0$ . Thus, in view of (1.2), for any  $u \in Q_2$  we have

$$(2.4) \quad d_2(u) \geq \frac{1}{2}\sigma_2\|u - u_0\|_2^2.$$

Let  $\mu_2$  be a positive *smoothness* parameter. Consider the following function:

$$(2.5) \quad f_{\mu_2}(x) = \hat{f}(x) + \max_{u \in Q_2} \{\langle Ax, u \rangle_2 - \hat{\phi}(u) - \mu_2 d_2(u)\}.$$

Denote by  $u_{\mu_2}(x)$  the optimal solution of above problem. Since function  $d_2(u)$  is strongly convex, this solution is unique. In accordance with Danskin's theorem, the gradient of  $f_{\mu_2}$  is well defined by

$$(2.6) \quad \nabla f_{\mu_2}(x) = \nabla \hat{f}(x) + A^* u_{\mu_2}(x).$$

Moreover, this gradient is Lipschitz-continuous with the constant

$$(2.7) \quad L_1(f_{\mu_2}) = L_1(\hat{f}) + \frac{1}{\sigma_2 \mu_2} \|A\|_{1,2}^2$$

(see, for example, Theorem 1 in [3]).

Similarly, let us consider a prox-function  $d_1(x)$  of the set  $Q_1$ , which has convexity parameter  $\sigma_1$  and the prox-center  $x_0$  with  $d_1(x_0) = 0$ . By (1.2), for any  $x \in Q_1$  we have

$$(2.8) \quad d_1(x) \geq \frac{1}{2}\sigma_1\|x - x_0\|_1^2.$$

Let  $\mu_1$  be a positive smoothness parameter. Consider

$$(2.9) \quad \phi_{\mu_1}(u) = -\hat{\phi}(u) + \min_{x \in Q_1} \{\langle Ax, u \rangle_2 + \hat{f}(x) + \mu_1 d_1(x)\}.$$

Since the second term in the above definition is a minimum of linear functions,  $\phi_{\mu_1}(u)$  is concave. Denote by  $x_{\mu_1}(u)$  the unique optimal solution of the above problem. In accordance with Danskin's theorem and Theorem 1 in [3], the gradient

$$(2.10) \quad \nabla \phi_{\mu_1}(u) = -\nabla \hat{\phi}(u) + Ax_{\mu_1}(u)$$

is Lipschitz-continuous with the constant

$$(2.11) \quad L_2(\phi_{\mu_1}) = L_2(\hat{\phi}) + \frac{1}{\sigma_1 \mu_1} \|A\|_{1,2}^2.$$

**3. Excessive gap condition.** Note that for any  $x \in Q_1$  and  $u \in Q_2$  we have

$$(3.1) \quad \phi(u) \leq f(x),$$

and our assumptions guarantee no duality gap for (2.1), (2.3). However,  $f_{\mu_2}(x) \leq f(x)$  and  $\phi(u) \leq \phi_{\mu_1}(u)$ . That opens up a possibility to satisfy the following *excessive gap condition*:

$$(3.2) \quad f_{\mu_2}(\bar{x}) \leq \phi_{\mu_1}(\bar{u})$$

by certain  $\bar{x} \in Q_1$  and  $\bar{u} \in Q_2$ . It is clear that (3.2) provides us with an upper bound on the quality of the primal-dual pair  $(\bar{x}, \bar{u})$ .

LEMMA 3.1. *Let  $\bar{x} \in Q_1$  and  $\bar{u} \in Q_2$  satisfy (3.2). Then*

$$(3.3) \quad 0 \leq \max\{f(\bar{x}) - f^*, f^* - \phi(\bar{u})\} \leq f(\bar{x}) - \phi(\bar{u}) \leq \mu_1 D_1 + \mu_2 D_2,$$

where  $D_1 = \max_{x \in Q_1} d_1(x)$ , and  $D_2 = \max_{u \in Q_2} d_2(u)$ .

*Proof.* Indeed, for any  $\bar{x} \in Q_1$ ,  $\bar{u} \in Q_2$  we have  $f(\bar{x}) - \mu_2 D_2 \leq f_{\mu_2}(\bar{x}) \leq \phi_{\mu_1}(\bar{u}) \leq \phi(\bar{u}) + \mu_1 D_1$ . It remains to apply (3.1).  $\square$

Our goal is to justify a process for updating recursively the pair  $(\bar{x}, \bar{u})$ , which keep satisfying inequality (3.2) as  $\mu_1$  and  $\mu_2$  go to zero. This can be done in two different ways, which correspond to two different auxiliary problems we must be ready to solve at each iteration. Before we start our analysis, let us prove one useful inequality.

LEMMA 3.2. *For any  $x$  and  $\bar{y}$  from  $Q_1$  we have:*

$$(3.4) \quad f_{\mu_2}(\bar{y}) + \langle \nabla f_{\mu_2}(\bar{y}), x - \bar{y} \rangle_1 \leq \hat{f}(x) + \langle Ax, u_{\mu_2}(\bar{y}) \rangle_2 - \hat{\phi}(u_{\mu_2}(\bar{y})).$$

*Proof.* Let us take arbitrary  $x$  and  $\bar{y}$  from  $Q_1$ . Denote  $\bar{u} = u_{\mu_2}(\bar{y})$ . Then

$$\begin{aligned} & f_{\mu_2}(\bar{y}) + \langle \nabla f_{\mu_2}(\bar{y}), x - \bar{y} \rangle_1 \\ & \text{(by (2.5), (2.6))} = \hat{f}(\bar{y}) + \langle A\bar{y}, \bar{u} \rangle_2 - \hat{\phi}(\bar{u}) - \mu_2 d_2(\bar{u}) + \langle \nabla \hat{f}(\bar{y}) + A^* \bar{u}, x - \bar{y} \rangle_1 \\ & \text{(by convexity of } \hat{f}) \leq \hat{f}(x) + \langle Ax, \bar{u} \rangle_2 - \hat{\phi}(\bar{u}). \end{aligned}$$

$\square$

**4. Gradient mapping.** Let us justify a possibility to satisfy the excessive gap condition (3.2) at some starting primal-dual pair. For  $x \in Q_1$  define the *primal gradient mapping*:

$$(4.1) \quad T_{\mu_2}(x) = \arg \min_{y \in Q_1} \left\{ \langle \nabla f_{\mu_2}(x), y - x \rangle_1 + \frac{1}{2} L_1(f_{\mu_2}) \|y - x\|_1^2 \right\}.$$

LEMMA 4.1. *Let us choose an arbitrary  $\mu_2 > 0$ . For prox-center  $x_0$  define*

$$(4.2) \quad \bar{x} = T_{\mu_2}(x_0), \quad \bar{u} = u_{\mu_2}(x_0).$$

*Then the excessive gap condition (3.2) is satisfied for any*

$$(4.3) \quad \mu_1 \geq \frac{1}{\sigma_1} L_1(f_{\mu_2}).$$

*Proof.* Denote  $\bar{x} = T_{\mu_2}(x_0)$ ,  $L_1 = L_1(f_{\mu_2})$ , and  $\bar{u} = u_{\mu_2}(x_0)$ . Since the gradient  $\nabla f_{\mu_2}$  is Lipschitz continuous, by (1.5) we have

$$\begin{aligned}
f_{\mu_2}(\bar{x}) &\leq f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_1 + \frac{1}{2} L_1 \|\bar{x} - x_0\|_1^2 \\
\text{(by (4.1))} &= \min_{x \in Q_1} \left\{ f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), x - x_0 \rangle_1 + \frac{1}{2} L_1 \|x - x_0\|_1^2 \right\} \\
\text{(by (3.4), (4.3))} &\leq \min_{x \in Q_1} \left\{ \hat{f}(x) + \langle Ax, \bar{u} \rangle_2 - \hat{\phi}(\bar{u}) + \frac{1}{2} \mu_1 \sigma_1 \|x - x_0\|_1^2 \right\} \\
\text{(by (2.8))} &\leq -\hat{\phi}(\bar{u}) + \min_{x \in Q_1} \left\{ \hat{f}(x) + \langle Ax, \bar{u} \rangle_2 + \mu_1 d_1(x) \right\} = \phi_{\mu_1}(\bar{u}).
\end{aligned}$$

□

Thus, the condition (3.2) can be satisfied for some primal-dual pair. Let us show how we can update the points  $\bar{x}$  and  $\bar{u}$  in order to keep (3.2) valid for smaller values of  $\mu_1$  and  $\mu_2$ . Note that in view of symmetry of the situation, at the first step of the process we can try to decrease only  $\mu_1$  keeping  $\mu_2$  unchanged. After that, at the second step, we update  $\mu_2$  and keep  $\mu_1$ , and so on. The main advantage of such a switching strategy is that we need to find a justification only for the first step. The proof for the second one will be symmetric.

**THEOREM 4.2.** *Let points  $\bar{x} \in Q_1$  and  $\bar{u} \in Q_2$  satisfy the excessive gap condition (3.2) for some positive  $\mu_1$  and  $\mu_2$ . Let us fix  $\tau \in (0, 1)$  and choose  $\mu_1^+ = (1 - \tau)\mu_1$ ,*

$$\begin{aligned}
\hat{x} &= (1 - \tau)\bar{x} + \tau x_{\mu_1}(\bar{u}), \\
\bar{u}_+ &= (1 - \tau)\bar{u} + \tau u_{\mu_2}(\hat{x}), \\
\bar{x}_+ &= T_{\mu_2}(\hat{x}).
\end{aligned}
\tag{4.4}$$

*Then the pair  $(\bar{x}_+, \bar{u}_+)$  satisfies condition (3.2) with smoothness parameters  $\mu_1^+$  and  $\mu_2$  provided that  $\tau$  is chosen in accordance to the following relation:*

$$\frac{\tau^2}{1 - \tau} \leq \frac{\mu_1 \sigma_1}{L_1(f_{\mu_2})}.
\tag{4.5}$$

*Proof.* Denote  $\hat{u} = u_{\mu_2}(\hat{x})$  and  $x_1 = x_{\mu_1}(\bar{u})$ . Since  $\hat{\phi}$  is convex, in view of the second line in (4.4) we have  $\hat{\phi}(\bar{u}_+) \leq (1 - \tau)\hat{\phi}(\bar{u}) + \tau\hat{\phi}(\hat{u})$ . Therefore

$$\begin{aligned}
\phi_{\mu_1^+}(\bar{u}_+) &= \min_{x \in Q_1} \left\{ (1 - \tau)\mu_1 d_1(x) + \langle Ax, (1 - \tau)\bar{u} + \tau\hat{u} \rangle_2 + \hat{f}(x) \right\} - \hat{\phi}(\bar{u}_+) \\
&\geq \min_{x \in Q_1} \left\{ (1 - \tau) \left[ \mu_1 d_1(x) + \langle Ax, \bar{u} \rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u}) \right] \right. \\
&\quad \left. + \tau [\hat{f}(x) + \langle Ax, \hat{u} \rangle_2 - \hat{\phi}(\hat{u})] \right\}.
\end{aligned}$$

Note that in view of condition (3.2) and the first line in (4.4) we have

$$\phi_{\mu_1}(\bar{u}) \geq f_{\mu_2}(\bar{x}) \geq f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - \hat{x} \rangle_1 = f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - x_1 \rangle_1.$$

Therefore, in view of property (1.2) and definition (2.9) we can estimate the expression in the first brackets as

follows:

$$\begin{aligned}
[\cdot]_1 &\geq \phi_{\mu_1}(\bar{u}) + \frac{1}{2}\mu_1\sigma_1\|x - x_1\|_1^2 \\
(\text{by (3.2)}) &\geq f_{\mu_2}(\bar{x}) + \frac{1}{2}\mu_1\sigma_1\|x - x_1\|_1^2 \\
(f \text{ is convex}) &\geq f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - \hat{x} \rangle_1 + \frac{1}{2}\mu_1\sigma_1\|x - x_1\|_1^2 \\
(\text{line 1, (4.4)}) &= f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - x_1 \rangle_1 + \frac{1}{2}\mu_1\sigma_1\|x - x_1\|_1^2.
\end{aligned}$$

In view of (3.4), for the second brackets we have

$$\begin{aligned}
[\cdot]_2 &\geq f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), x - \hat{x} \rangle_1 \\
(\text{line 1, (4.4)}) &= f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), x - x_1 + (1 - \tau)(x_1 - \bar{x}) \rangle_1.
\end{aligned}$$

Thus, putting all together, we accomplish the proof as follows:

$$\begin{aligned}
\phi_{\mu_1}^+(\bar{u}_+) &\geq \min_{x \in Q_1} \left\{ f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), x - x_1 \rangle_1 + \frac{1}{2}(1 - \tau)\mu_1\sigma_1\|x - x_1\|_1^2 \right\} \\
(\text{by (4.5)}) &\geq \min_{x \in Q_1} \left\{ f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), x - x_1 \rangle_1 + \frac{1}{2}\tau^2 L_1(f_{\mu_2})\|x - x_1\|_1^2 \right\} \\
(y = \bar{x} + \tau(x - \bar{x})) &= \min_{y \in \bar{x} + \tau(Q_1 - \bar{x})} \left\{ f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), y - \hat{x} \rangle_1 + \frac{1}{2}L_1(f_{\mu_2})\|y - \hat{x}\|_1^2 \right\} \\
(Q_1 \text{ is convex}) &\geq \min_{y \in Q_1} \left\{ f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), y - \hat{x} \rangle_1 + \frac{1}{2}L_1(f_{\mu_2})\|y - \hat{x}\|_1^2 \right\} \\
(\text{line 3, (4.4)}) &= f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x}_+ - \hat{x} \rangle_1 + \frac{1}{2}L_1(f_{\mu_2})\|\bar{x}_+ - \hat{x}\|_1^2 \\
(\text{by (1.5)}) &\geq f_{\mu_2}(\bar{x}_+).
\end{aligned}$$

□

**5. Bregman projection.** Let us assume for simplicity that  $d_1(x)$  is differentiable. Then for any  $x \in Q_1$  we have

$$(5.1) \quad \langle \nabla d_1(x_0), x - x_0 \rangle_1 \geq 0.$$

For  $x$  and  $z$  from  $Q_1$  denote by

$$\xi_1(z, x) = d_1(x) - d_1(z) - \langle \nabla d_1(z), x - z \rangle_1$$

the *Bregman distance* between  $z$  and  $x$ . If  $z$  is fixed, then  $\xi(z, x)$  is strongly convex in  $x$ . Moreover, in view of (1.3)

$$(5.2) \quad \xi_1(z, x) \geq \frac{1}{2}\sigma_1\|x - z\|_1^2.$$

Define the *Bregman projection* of some  $g \in E_1^*$  onto the set  $Q_1$  as follows:

$$(5.3) \quad V_1(z, g) = \arg \min_{x \in Q_1} \{ \langle g, x - z \rangle_1 + \xi_1(z, x) \}.$$

As compared with the gradient mapping, Bregman projection has several advantages. Firstly, it is uniquely defined. Secondly, the optimization problem in (5.3) involves the same objects as (2.9). So, we have more chances to have it solvable in an easy way (see Section 5.3 in [3] for an example).

Let us show that the Bregman projection can be also used for finding a primal-dual pair, which satisfies the excessive gap condition (3.2).

LEMMA 5.1. *Let us choose an arbitrary  $\mu_2 > 0$ . Denote  $\gamma = \frac{\sigma_1}{L_1(f_{\mu_2})}$  and set*

$$(5.4) \quad \bar{x} = V_1(x_0, \gamma \nabla f_{\mu_2}(x_0)), \quad \bar{u} = u_{\mu_2}(x_0).$$

*Then the excessive gap condition is satisfied for any  $\mu_1 \geq \gamma^{-1}$ .*

*Proof.* Indeed, in view of (1.5) we have

$$\begin{aligned} f_{\mu_2}(\bar{x}) &\leq f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_1 + \frac{1}{2} L_1(f_{\mu_2}) \|\bar{x} - x_0\|_1^2 \\ &= f_{\mu_2}(x_0) + \frac{1}{\gamma} [\langle \gamma \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_1 + \frac{1}{2} \sigma_1 \|\bar{x} - x_0\|_1^2] \\ \text{(by (5.2))} &\leq f_{\mu_2}(x_0) + \frac{1}{\gamma} [\langle \gamma \nabla f_{\mu_2}(x_0), \bar{x} - x_0 \rangle_1 + \xi_1(x_0, \bar{x})] \\ \text{(by (5.3), (5.4))} &= f_{\mu_2}(x_0) + \frac{1}{\gamma} \min_{x \in Q_1} \{ \langle \gamma \nabla f_{\mu_2}(x_0), x - x_0 \rangle_1 + \xi_1(x_0, x) \} \\ &= \min_{x \in Q_1} \left\{ f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), x - x_0 \rangle_1 + \frac{1}{\gamma} \xi_1(x_0, x) \right\} \\ \text{(by (5.1))} &\leq \min_{x \in Q_1} \left\{ f_{\mu_2}(x_0) + \langle \nabla f_{\mu_2}(x_0), x - x_0 \rangle_1 + \frac{1}{\gamma} d_1(x) \right\} \\ \text{(using (3.4))} &\leq \min_{x \in Q_1} \left\{ \hat{f}(x) + \langle Ax, u_{\mu_2}(x_0) \rangle - \hat{\phi}(u_{\mu_2}(x_0)) + \frac{1}{\gamma} d_1(x) \right\} \\ \text{(by (2.9))} &= \phi_{\gamma^{-1}}(u_{\mu_2}(x_0)) \leq \phi_{\mu_1}(u_{\mu_2}(x_0)). \end{aligned}$$

□

As in Section 4, we present a justification only for the first (primal) step of the switching primal-dual strategy for maintaining the excessive gap condition (3.2) while the parameters  $\mu_1$  and  $\mu_2$  go to zero.

THEOREM 5.2. *Let points  $\bar{x} \in Q_1$  and  $\bar{u} \in Q_2$  satisfy the excessive gap condition (3.2) for some positive  $\mu_1$  and  $\mu_2$ . Let us choose  $\tau \in (0, 1)$  in accordance with (4.5) and set*

$$(5.5) \quad \begin{aligned} \hat{x} &= (1 - \tau)\bar{x} + \tau x_{\mu_1}(\bar{u}), \\ \bar{u}_+ &= (1 - \tau)\bar{u} + \tau u_{\mu_2}(\hat{x}), \\ \tilde{x} &= V_1(x_{\mu_1}(\bar{u}), \frac{\tau}{(1-\tau)\mu_1} \nabla f_{\mu_2}(\hat{x})), \\ \bar{x}_+ &= (1 - \tau)\bar{x} + \tau \tilde{x}, \\ \mu_1^+ &= (1 - \tau)\mu_1. \end{aligned}$$

*Then the pair  $(\bar{x}_+, \bar{u}_+)$  satisfies (3.2) with the smoothness parameters  $\mu_1^+$  and  $\mu_2$ .*

*Proof.* Denote  $\hat{u} = u_{\mu_2}(\hat{x})$  and  $x_1 = x_{\mu_1}(\bar{u})$ . In view of the rules (5.5), convexity of  $\hat{\phi}$ , and inequality (3.4), we have:

$$\begin{aligned} &(1 - \tau)\mu_1 d_1(x) + \langle Ax, (1 - \tau)\bar{u} + \tau \hat{u} \rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u}_+) \\ &\geq (1 - \tau) \left[ \mu_1 d_1(x) + \langle Ax, \bar{u} \rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u}) \right] + \tau [\hat{f}(x) + \langle Ax, \hat{u} \rangle_2 - \hat{\phi}(\hat{u})] \\ &\geq (1 - \tau) \left[ \mu_1 d_1(x) + \langle Ax, \bar{u} \rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u}) \right]_1 + \tau [f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), x - \hat{x} \rangle_1]_2. \end{aligned}$$

The first order optimality conditions for point  $x_1$  are as follows:

$$(5.6) \quad \langle \mu_1 \nabla d_1(x_1) + A^* \bar{u} + \nabla \hat{f}(x_1), x - x_1 \rangle_1 \geq 0, \quad x \in Q_1.$$

Therefore, using convexity of  $\hat{f}$  and  $f_{\mu_2}$ , we can estimate the term  $[\cdot]_1$  as follows:

$$\begin{aligned} [\cdot]_1 &= \mu_1 (\xi(x_1, x) + d_1(x_1) + \langle \nabla d_1(x_1), x - x_1 \rangle_1) + \langle Ax, \bar{u} \rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u}) \\ (\text{by (5.6)}) &\geq \mu_1 \xi(x_1, x) + \mu_1 d_1(x_1) + \langle Ax_1, \bar{u} \rangle_2 + \hat{f}(x) - \langle \nabla \hat{f}(x_1), x - x_1 \rangle_1 - \hat{\phi}(\bar{u}) \\ &\geq \mu_1 \xi(x_1, x) + \mu_1 d_1(x_1) + \langle Ax_1, \bar{u} \rangle_2 + \hat{f}(x_1) - \hat{\phi}(\bar{u}) \\ (\text{by (2.9)}) &= \mu_1 \xi(x_1, x) + \phi_{\mu_1}(\bar{u}) \\ (\text{by (3.2)}) &\geq \mu_1 \xi(x_1, x) + f_{\mu_2}(\bar{x}) \\ &\geq \mu_1 \xi(x_1, x) + f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x} - \hat{x} \rangle_1. \end{aligned}$$

Thus, we can continue:

$$\begin{aligned} \phi_{\mu_1^+}(\bar{u}_+) &= \min_{x \in Q_1} \left\{ (1 - \tau) \mu_1 d_1(x) + \langle Ax, (1 - \tau) \bar{u} + \tau \hat{u} \rangle_2 + \hat{f}(x) \right\} - \hat{\phi}(\bar{u}_+) \\ &\geq \min_{x \in Q_1} \left\{ (1 - \tau) \mu_1 \xi(x_1, x) + f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), (1 - \tau) \bar{x} + \tau x - \hat{x} \rangle_1 \right\} \\ (\text{line 1, (5.5)}) &= \min_{x \in Q_1} \left\{ (1 - \tau) \mu_1 \xi(x_1, x) + f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), x - x_1 \rangle_1 \right\} \\ (\text{line 3, (5.5)}) &= (1 - \tau) \mu_1 \xi(x_1, \tilde{x}) + f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), \tilde{x} - x_1 \rangle_1 \\ (\text{by (5.2)}) &\geq \frac{1}{2} (1 - \tau) \mu_1 \sigma_1 \|\tilde{x} - x_1\|_1^2 + f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), \tilde{x} - x_1 \rangle_1 \\ (\text{by (4.5)}) &\geq \frac{1}{2} \tau^2 L_1(f_{\mu_2}) \|\tilde{x} - x_1\|_1^2 + f_{\mu_2}(\hat{x}) + \tau \langle \nabla f_{\mu_2}(\hat{x}), \tilde{x} - x_1 \rangle_1 \\ (\text{line 4, (5.5)}) &= \frac{1}{2} L_1(f_{\mu_2}) \|\bar{x}_+ - \hat{x}\|_1^2 + f_{\mu_2}(\hat{x}) + \langle \nabla f_{\mu_2}(\hat{x}), \bar{x}_+ - \hat{x} \rangle_1 \\ (\text{by (1.5)}) &\geq f_{\mu_2}(\bar{x}_+). \end{aligned}$$

□

**6. Convergence analysis.** In Sections 4, 5 we have seen that the smoothness parameters  $\mu_1$  and  $\mu_2$  can be decreased by a switching strategy. Thus, in order to convert the results of Theorems 4.2, 5.2 in an algorithmic scheme we only need to point out a strategy for updating these parameters, which is compatible with the condition (4.5). In this section we do that for an important case  $L_1(\hat{f}) = L_2(\hat{\phi}) = 0$ .

It is convenient to represent the smoothness parameters as follows:

$$(6.1) \quad \mu_1 = \lambda_1 \cdot \|A\|_{1,2} \cdot \sqrt{\frac{D_2}{\sigma_1 \sigma_2 D_1}}, \quad \mu_2 = \lambda_2 \cdot \|A\|_{1,2} \cdot \sqrt{\frac{D_1}{\sigma_1 \sigma_2 D_2}}.$$

Then the estimate (3.3) for the duality gap becomes symmetric:

$$(6.2) \quad f(\bar{x}) - \phi(\bar{u}) \leq (\lambda_1 + \lambda_2) \cdot \|A\|_{1,2} \cdot \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}}.$$

Since by (2.7),  $L_1(f_{\mu_2}) = \frac{1}{\sigma_2 \mu_2} \|A\|_{1,2}^2$ , the condition (4.5) becomes problem independent:

$$(6.3) \quad \frac{\tau^2}{1 - \tau} \leq \mu_1 \mu_2 \cdot \frac{\sigma_1 \sigma_2}{\|A\|_{1,2}^2} = \lambda_1 \lambda_2.$$

Let us write down the switching algorithmic scheme in an explicit form. It is convenient to have a permanent iteration counter. In this case at even iterations we apply the primal update (4.4) (or (5.5)), and at odd iterations we apply the corresponding dual update. Since at even iterations  $\lambda_2$  is not changing, and at odd iterations  $\lambda_1$  is not changing, it is convenient to put their new values in same sequence  $\{\alpha_k\}_{k=-1}^{\infty}$ . Let us fix the following relations between the sequences:

$$(6.4) \quad \begin{aligned} k = 2l & : \lambda_{1,k} = \alpha_{k-1}, \quad \lambda_{2,k} = \alpha_k, \\ k = 2l + 1 & : \lambda_{1,k} = \alpha_k, \quad \lambda_{2,k} = \alpha_{k-1}. \end{aligned}$$

Then the parameters  $\tau_k$  define the reduction rate of the sequence  $\{\alpha_k\}_{k=-1}^{\infty}$ .

LEMMA 6.1. *For all  $k \geq 0$  we have  $\alpha_{k+1} = (1 - \tau_k)\alpha_{k-1}$ .*

*Proof.* Indeed, in accordance to (6.4), if  $k = 2l$ , then

$$\alpha_{k+1} = \lambda_{1,k+1} = (1 - \tau_k)\lambda_{1,k} = (1 - \tau_k)\alpha_{k-1}.$$

And if  $k = 2l + 1$ , then  $\alpha_{k+1} = \lambda_{2,k+1} = (1 - \tau_k)\lambda_{2,k} = (1 - \tau_k)\alpha_{k-1}$ . □

COROLLARY 6.2. *In terms of the sequence  $\{\alpha_k\}_{k=-1}^{\infty}$  the condition (6.3) looks as follows:*

$$(6.5) \quad (\alpha_{k+1} - \alpha_{k-1})^2 \leq \alpha_{k+1}\alpha_k\alpha_{k-1}^2, \quad k \geq 0.$$

*Proof.* In view of (6.4) we always have  $\lambda_{1,k}\lambda_{2,k} = \alpha_k\alpha_{k-1}$ . Since  $\tau_k = 1 - \frac{\alpha_{k+1}}{\alpha_{k-1}}$ , we get (6.5). □

Clearly, condition (6.5) is satisfied by

$$(6.6) \quad \alpha_k = \frac{2}{k+2}, \quad k \geq -1.$$

Then

$$(6.7) \quad \tau_k = 1 - \frac{\alpha_{k+1}}{\alpha_{k-1}} = \frac{2}{k+3}, \quad k \geq 0.$$

Now we are ready to write down the algorithmic scheme. Let us do that for the gradient mapping update (4.4). In this scheme we use the sequences  $\{\mu_{1,k}\}_{k=-1}^{\infty}$  and  $\{\mu_{2,k}\}_{k=-1}^{\infty}$ , generated in accordance to rules (6.1), (6.4) and (6.6).

**1. Initialization:**

Choose  $\bar{x}_0$  and  $\bar{u}_0$  in accordance to (4.2) with  $\mu_1 = \mu_{1,0}$  and  $\mu_2 = \mu_{2,0}$ .

**2. Iterations ( $k \geq 0$ ):**

- (6.8) a) Set  $\tau_k = \frac{2}{k+3}$ .
- b) If  $k$  is even, then generate  $(\bar{x}_{k+1}, \bar{u}_{k+1})$  from  $(\bar{x}_k, \bar{u}_k)$  using (4.4).
- c) If  $k$  is odd, then generate  $(\bar{x}_{k+1}, \bar{u}_{k+1})$  from  $(\bar{x}_k, \bar{u}_k)$  using the symmetric dual variant of (4.4).

THEOREM 6.3. *Let the sequences  $\{\bar{x}_k\}_{k=0}^{\infty}$  and  $\{\bar{u}_k\}_{k=0}^{\infty}$  be generated by method (6.8). Then each pair of points satisfy the excessive gap condition. Therefore*

$$(6.9) \quad f(\bar{x}_k) - \phi(\bar{u}_k) \leq \frac{4\|A\|_{1,2}}{k+1} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}}.$$

*Proof.* In accordance to our choice of parameters,

$$\mu_{1,0}\mu_{2,0} = \lambda_{1,0}\lambda_{2,0} \cdot \frac{\|A\|_{1,2}^2}{\sigma_1 \sigma_2} = \frac{2\mu_{2,0}}{\sigma_1} L_1(f_{\mu_{2,0}}) > \frac{\mu_{2,0}}{\sigma_1} L_1(f_{\mu_{2,0}}).$$

Hence, in view of Lemma 4.1 the pair  $(\bar{x}_0, \bar{u}_0)$  satisfies the excessive gap condition. We have already checked that the sequence  $\{\tau_k\}_{k=0}^{\infty}$  defined by (6.7) satisfies the conditions of Theorem 4.2. Therefore the excessive gap conditions will be valid for the sequences generated by (6.8). It remains to use inequality (6.2).  $\square$

Clearly, the same statement is valid for the method based on the updating scheme (5.5).

**7. Minimizing a strongly convex function.** Consider now the model (2.2), which satisfies the following assumption.

ASSUMPTION 1. *In representation (2.2) function  $\hat{f}(x)$  is strongly convex with strong convexity parameter  $\hat{\sigma} > 0$ .*

Let us prove the following variant of Danskin's theorem.

LEMMA 7.1. *Under Assumption 1, function  $\phi(u)$  defined by (2.3) is concave and differentiable. Moreover, its gradient*

$$(7.1) \quad \nabla\phi(u) = -\nabla\hat{\phi}(u) + Ax_0(u),$$

with  $x_0(u)$  defined by (2.9), is Lipschitz-continuous with the constant

$$(7.2) \quad L_2(\phi) = \frac{1}{\hat{\sigma}}\|A\|_{1,2}^2 + L_2(\hat{\phi}).$$

*Proof.* Denote  $\tilde{\phi}(u) = \min_{x \in Q_1} \{\langle Ax, u \rangle_2 + \hat{f}(x)\}$ . This function is concave as a minimum of linear functions. Since  $\hat{f}$  is strongly convex, the solution of the above minimization problem is unique. Therefore  $\tilde{\phi}(u)$  is differentiable and  $\nabla\tilde{\phi}(u) = Ax_0(u)$ .

Consider two points  $u_1$  and  $u_2$ . From the first-order optimality conditions for (2.3) we have

$$\langle A^*u_1 + \nabla\hat{f}(x_0(u_1)), x_0(u_2) - x_0(u_1) \rangle_1 \geq 0,$$

$$\langle A^*u_2 + \nabla\hat{f}(x_0(u_2)), x_0(u_1) - x_0(u_2) \rangle_1 \geq 0.$$

Adding these inequalities and using strong convexity of  $\hat{f}(\cdot)$ , we continue as follows:

$$\begin{aligned} \langle Ax_0(u_2) - Ax_0(u_1), u_1 - u_2 \rangle_2 &\geq \langle \nabla\hat{f}(x_0(u_1)) - \nabla\hat{f}(x_0(u_2)), x_0(u_1) - x_0(u_2) \rangle_1 \\ &\text{(by (1.4))} \geq \hat{\sigma}\|x_0(u_1) - x_0(u_2)\|_1^2 \\ &\text{(by (1.1))} \geq \frac{\hat{\sigma}}{\|A\|_{1,2}^2} \left( \|\nabla\tilde{\phi}(u_1) - \nabla\tilde{\phi}(u_2)\|_2^* \right)^2. \end{aligned}$$

Thus,  $\|\nabla\tilde{\phi}(u_1) - \nabla\tilde{\phi}(u_2)\|_2^* \leq \frac{1}{\hat{\sigma}}\|A\|_{1,2}^2 \cdot \|u_1 - u_2\|_2$ , and (7.2) follows.  $\square$

LEMMA 7.2. *For any  $u$  and  $\hat{u}$  from  $Q_2$  we have:*

$$(7.3) \quad \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u} \rangle_2 \geq -\hat{\phi}(u) + \langle Ax_0(\hat{u}), u \rangle_2 + \hat{f}(x_0(\hat{u})).$$

*Proof.* Let us take arbitrary  $u$  and  $\hat{u}$  from  $Q_2$ . Denote  $\hat{x} = x_0(\hat{u})$ . Then

$$\begin{aligned} \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u} \rangle_2 &= -\hat{\phi}(\hat{u}) + \langle A\hat{x}, \hat{u} \rangle_2 + \hat{f}(\hat{x}) + \langle -\nabla\hat{\phi}(\hat{u}) + A\hat{x}, u - \hat{u} \rangle_2 \\ &\text{(\hat{\phi} is convex)} \geq -\hat{\phi}(u) + \langle A\hat{x}, u \rangle_2 + \hat{f}(\hat{x}). \end{aligned}$$

$\square$

In this section we derive an optimization scheme from the following variant of the excessive gap condition:

$$(7.4) \quad f_{\mu_2}(\bar{x}) \leq \phi(\bar{u})$$

for some  $\bar{x} \in Q_1$  and  $\bar{u}$  in  $Q_2$ .

This condition can be seen as a variant of condition (3.2) with  $\mu_1 = 0$ . However, in this section we prefer not to use the results of the previous sections since our assumptions are now different. For example, we do not need anymore the set  $Q_1$  to be bounded.

LEMMA 7.3. *Let points  $\bar{x}$  from  $Q_1$  and  $\bar{u}$  from  $Q_2$  satisfy (7.4). Then*

$$(7.5) \quad 0 \leq f(\bar{x}) - \phi(\bar{u}) \leq \mu_2 D_2.$$

*Proof.* Indeed, for any  $x \in Q_1$  we have  $f_{\mu_2}(x) \geq f(x) - \mu_2 D_2$ . □

Define the adjoint gradient mapping as follows:

$$(7.6) \quad V(u) = \arg \max_{v \in Q_2} \left\{ \langle \nabla \phi(u), v - u \rangle_2 - \frac{1}{2} L_2(\phi) \|v - u\|_2^2 \right\}.$$

LEMMA 7.4. *The excessive gap condition (7.4) is valid for  $\mu_2 = \frac{1}{\sigma_2} L_2(\phi)$  and*

$$(7.7) \quad \bar{x} = x_0(u_0), \quad \bar{u} = V(u_0).$$

*Proof.* Indeed, in view of Lemma 7.1 and (1.5) we get the following relations:

$$\begin{aligned} \phi(V(u_0)) &\geq \phi(u_0) + \langle \nabla \phi(u_0), V(u_0) - u_0 \rangle_2 - \frac{1}{2} L_2(\phi) \|V(u_0) - u_0\|_2^2 \\ &\text{(by (7.6))} = \max_{u \in Q_2} \left\{ \phi(u_0) + \langle \nabla \phi(u_0), u - u_0 \rangle_2 - \frac{1}{2} L_2(\phi) \|u - u_0\|_2^2 \right\} \\ &\text{(by (2.3) and (7.1))} = \max_{u \in Q_2} \left\{ -\hat{\phi}(u_0) + \langle Ax_0(u_0), u_0 \rangle_2 + \hat{f}(x_0(u_0)) \right. \\ &\quad \left. + \langle Ax_0(u_0) - \nabla \hat{\phi}(u_0), u - u_0 \rangle_2 - \frac{1}{2} \mu_2 \sigma_2 \|u - u_0\|_2^2 \right\} \\ &\text{(\hat{\phi} is convex and (2.4))} \geq \max_{u \in Q_2} \left\{ -\hat{\phi}(u) + \hat{f}(x_0(u_0)) + \langle Ax_0(u_0), u \rangle_2 - \mu_2 d_2(u) \right\} \\ &\text{(by (2.5))} = f_{\mu_2}(x_0(u_0)). \end{aligned}$$

□

THEOREM 7.5. *Let points  $\bar{x} \in Q_1$  and  $\bar{u} \in Q_2$  satisfy the excessive gap condition (7.4) for some positive  $\mu_2$ . Let us fix  $\tau \in (0, 1)$  and choose  $\mu_2^+ = (1 - \tau)\mu_2$ ,*

$$(7.8) \quad \begin{aligned} \hat{u} &= (1 - \tau)\bar{u} + \tau u_{\mu_2}(\bar{x}), \\ \bar{x}_+ &= (1 - \tau)\bar{x} + \tau x_0(\hat{u}), \\ \bar{u}_+ &= V(\hat{u}). \end{aligned}$$

*Then the pair  $(\bar{x}_+, \bar{u}_+)$  satisfies condition (7.4) with smoothness parameter  $\mu_2$ , provided that  $\tau$  is chosen in accordance to the following relation:*

$$(7.9) \quad \frac{\tau^2}{1 - \tau} \leq \frac{\mu_2 \sigma_2}{L_2(\phi)}.$$

*Proof.* Denote  $\hat{x} = x_0(\hat{u})$  and  $u_2 = u_{\mu_2}(\bar{x})$ . In view of second line in (7.8) and (2.5) we have:

$$\begin{aligned}
f_{\mu_2^+}(\bar{x}_+) &= \hat{f}(\bar{x}_+) + \max_{u \in Q_2} \left\{ \langle A((1-\tau)\bar{x} + \tau\hat{x}), u \rangle_2 - \hat{\phi}(u) - (1-\tau)\mu_2 d_2(u) \right\} \\
(\hat{f} \text{ is convex}) &\leq \max_{u \in Q_2} \left\{ (1-\tau) \left[ \hat{f}(\bar{x}) + \langle A\bar{x}, u \rangle_2 - \hat{\phi}(u) - \mu_2 d_2(u) \right] \right. \\
&\quad \left. + \tau \left[ \hat{f}(\hat{x}) + \langle A\hat{x}, u \rangle_2 - \hat{\phi}(u) \right] \right\} \\
(\text{by (1.2)}) &\leq \max_{u \in Q_2} \left\{ (1-\tau) \left[ f_{\mu_2}(\bar{x}) - \frac{1}{2}\mu_2\sigma_2 \|u - u_2\|_2^2 \right] \right. \\
(\text{by (7.3)}) &\quad \left. + \tau \left[ \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), u - \hat{u} \rangle_2 \right] \right\}.
\end{aligned}$$

Since  $\phi$  is concave, by (7.4) we obtain

$$\begin{aligned}
f_{\mu_2}(\bar{x}) &\leq \phi(\bar{u}) \leq \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), \bar{u} - \hat{u} \rangle_2 \\
(\text{line 1, (7.8)}) &= \phi(\hat{u}) + \tau \langle \nabla\phi(\hat{u}), \bar{u} - u_2 \rangle_2.
\end{aligned}$$

Hence, we can finish the proof as follows:

$$\begin{aligned}
f_{\mu_2^+}(\bar{x}_+) &\leq \max_{u \in Q_2} \left\{ \phi(\hat{u}) + \tau \langle \nabla\phi(\hat{u}), u - u_2 \rangle_2 - \frac{1}{2}(1-\tau)\mu_2\sigma_2 \|u - u_2\|_2^2 \right\} \\
(\text{by (7.9)}) &\leq \max_{u \in Q_2} \left\{ \phi(\hat{u}) + \tau \langle \nabla\phi(\hat{u}), u - u_2 \rangle_2 - \frac{1}{2}\tau^2 L_2(\phi) \|u - u_2\|_2^2 \right\} \\
(v = \bar{u} + \tau(u - \bar{u})) &= \max_{v \in \bar{u} + \tau(Q_2 - \bar{u})} \left\{ \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), v - \hat{u} \rangle_2 - \frac{1}{2}L_2(\phi) \|v - \hat{u}\|_2^2 \right\} \\
(Q_2 \text{ is convex}) &\leq \max_{v \in Q_2} \left\{ \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), v - \hat{u} \rangle_2 - \frac{1}{2}L_2(\phi) \|v - \hat{u}\|_2^2 \right\} \\
(\text{by (7.6)}) &= \phi(\hat{u}) + \langle \nabla\phi(\hat{u}), \bar{u}_+ - \hat{u} \rangle_2 - \frac{1}{2}L_2(\phi) \|\bar{u}_+ - \hat{u}\|_2^2 \\
(\text{by (1.5)}) &\leq \phi(\bar{u}_+).
\end{aligned}$$

□

Now we can justify the following minimization scheme.

**1. Initialization:**

Set  $\mu_{2,0} = \frac{2}{\sigma_2} L_2(\phi)$ ,  $\bar{x}_0 = x_0(u_0)$  and  $\bar{u}_0 = V(u_0)$ .

**2. For  $k \geq 0$  iterate :**

$$(7.10) \quad \text{Set } \tau_k = \frac{2}{k+3} \text{ and } \hat{u}_k = (1 - \tau_k)\bar{u}_k + \tau_k u_{\mu_{2,k}}(\bar{x}_k).$$

$$\text{Update } \mu_{2,k+1} = (1 - \tau_k)\mu_{2,k},$$

$$\bar{x}_{k+1} = (1 - \tau_k)\bar{x}_k + \tau_k x_0(\hat{u}_k),$$

$$\bar{u}_{k+1} = V(\hat{u}_k).$$

THEOREM 7.6. *Let problem (2.1) satisfy Assumption 1. Then the pairs  $(\bar{x}_k, \bar{u}_k)$  generated by scheme (7.10) satisfy the following inequality:*

$$(7.11) \quad f(\bar{x}_k) - \phi(\bar{u}_k) \leq \frac{4L_2(\phi)D_2}{(k+1)(k+2)\sigma_2},$$

where  $L_2(\phi)$  is given by (7.2).

*Proof.* Indeed, in view of Theorem 7.5 and Lemma 7.4 we need only to justify that the sequences  $\{\mu_{2,k}\}_{k=0}^\infty$  and  $\{\tau_k\}_{k=0}^\infty$  satisfy relation (7.9). That is straightforward because of relation

$$\mu_{2,k} = \frac{4L_2(\phi)}{(k+1)(k+2)\sigma_2},$$

which is valid for all  $k \geq 0$ . □

Let us conclude the paper with an example. Consider the problem

$$(7.12) \quad f(x) = \frac{1}{2}\|x\|_1^2 + \max_{1 \leq j \leq m} [f_j + \langle g_j, x - x_j \rangle_1] \rightarrow \min : x \in E_1.$$

The problems of this type arise, for example, at each iteration of Bundle Method [1]. Let  $E_1 = R^n$  and we choose

$$\|x\|_1^2 = \sum_{i=1}^n (x^{(i)})^2, \quad x \in E_1.$$

Then this problem can be solved by the scheme (7.10).

Indeed, we can represent the objective function in (7.12) in the form (2.2) using the following objects:

$$E_2 = R^m, \quad Q_2 = \Delta_m = \{u \in R_+^m : \sum_{j=1}^m u^{(j)} = 1\},$$

$$\hat{f}(x) = \frac{1}{2}\|x\|_1^2, \quad \hat{\phi}(u) = \langle b, u \rangle_2, \quad b^{(j)} = \langle g_j, x_j \rangle_1 - f_j, \quad j = 1, \dots, m,$$

$$A^T = (a_1, \dots, a_m).$$

Thus,  $\hat{\sigma} = 1$  and  $L_2(\hat{\phi}) = 0$ . Let us choose for  $E_2$  the following norm:

$$\|u\|_2 = \sum_{j=1}^m |u^{(j)}|.$$

Then we can use the entropy distance function (see [3]):

$$d_2(u) = \ln m + \sum_{j=1}^m u^{(j)} \ln u^{(j)}, \quad u_0 = \left(\frac{1}{m}, \dots, \frac{1}{m}\right),$$

for which  $\sigma_2 = 1$  and  $D_2 = \ln m$ . Note that in this case

$$\|A\|_{1,2} = \max_{1 \leq j \leq m} \|g_j\|_1^*.$$

Thus, method (7.10) as applied to the problem (7.12) converges with the following rate:

$$f(\bar{x}_k) - \phi(\bar{u}_k) \leq \frac{4 \ln m}{(k+1)(k+2)} \cdot \max_{1 \leq j \leq m} (\|g_j\|_1^*)^2.$$

Let us study the complexity of this scheme for our example. At each iteration we need to compute the following objects.

1. **Computation of  $u_{\mu_2}(\bar{x})$ .** This is the solution of the following problem:

$$\max_u \left\{ \sum_{j=1}^m u^{(j)} s^{(j)}(\bar{x}) - \mu_2 d_2(u) : u \in Q_2 \right\}$$

with  $s^{(j)}(\bar{x}) = f_j + \langle g_j, \bar{x} - x_j \rangle$ ,  $j = 1, \dots, m$ . In accordance to (4.14) in Lemma 4 [3], this solution can be found in a closed form:

$$u_{\mu_2}^{(j)}(\bar{x}) = e^{s^{(j)}(\bar{x})/\mu_2} \cdot \left[ \sum_{l=1}^m e^{s^{(l)}(\bar{x})/\mu_2} \right]^{-1}, \quad j = 1, \dots, m.$$

2. **Computation of  $x_0(\hat{u})$ .** In our case this is a solution to the problem

$$\min_x \{ \langle Ax, \hat{u} \rangle_2 + \frac{1}{2} \|x\|_1^2 : x \in E_1 \}.$$

Hence, the answer is very simple:  $x_0(\hat{u}) = -A^T \hat{u}$ .

3. **Computation of  $V(\hat{u})$ .** In our case

$$\begin{aligned} \phi(\bar{u}) &= \min_{x \in E_1} \left\{ \sum_{j=1}^m u^{(j)} [f_j + \langle g_j, x - x_j \rangle_1] + \frac{1}{2} \|x\|_1^2 \right\} \\ &= -\langle b, u \rangle_2 - \frac{1}{2} (\|A^T \hat{u}\|_1^*)^2. \end{aligned}$$

Thus,  $\nabla \phi(\bar{u}) = -b - AA^T \hat{u}$ . Now we can compute  $V(\hat{u})$  by (7.6). In [3], Section 5.1, it is shown that the complexity of finding  $V(\bar{u})$  is of the order  $O(m \ln m)$ .

We have seen that all computations at each iteration of the method (7.10) as applied to the problem (7.12) are very cheap. The most expensive part of the iteration is the multiplication of the matrix  $A$  by a vector. In a straightforward implementation we need three such multiplications per iteration. However, a simple modification of the order of operations can reduce this amount up to two.

**Acknowledgments.** The author thanks the anonymous referees for their very useful suggestions.

#### REFERENCES

- [1] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
- [2] A. NEMIROVSKY AND D. YUDIN, *Informational Complexity and Efficient Methods for Solution of Convex Extremal Problems*, J. Wiley & Sons, New York, 1983.
- [3] YU. NESTEROV, *Smooth minimization of non-smooth functions.*, CORE Discussion Paper #2003/12, January 2003, accepted by Mathematical Programming.
- [4] YU. NESTEROV, *Introductory lectures on Convex Optimization: Basic course*, Kluwer, Boston, 2004.