SIAM J. OPTIM.
Vol. 16, No. 1, pp. 235–249

EXCESSIVE GAP TECHNIQUE IN NONSMOOTH CONVEX MINIMIZATION*

YU. NESTEROV†

Abstract. In this paper we introduce a new primal-dual technique for convergence analysis of gradient schemes for nonsmooth convex optimization. As an example of its application, we derive a primal-dual gradient method for a special class of structured nonsmooth optimization problems, which ensures a rate of convergence of order $O(\frac{1}{k})$, where k is the iteration count. Another example is a gradient scheme, which minimizes a nonsmooth strongly convex function with known structure with 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, nonsmooth optimization, complexity theory, black-box oracle, optimal methods, structural optimization

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

DOI. 10.1137/S1052623403422285

1. Introduction.

This paper continues the research started in [3], where it was shown that some structured nonsmooth optimization problems can be solved with efficiency estimates $O(\frac{1}{\epsilon})$, where ϵ 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 a certain relaxation of the standard black-box assumption: it was assumed that our problem had an explicit and quite simple minimax structure. The numerical scheme proposed in [3] had a drawback, which decreases its practical efficiency: the number of steps must be fixed in advance, chosen in accordance with a worst-case complexity analysis.

In this paper we propose several new primal-dual gradient schemes for the same class of problems as those 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 sections 4 and 5 we present two different strategies for maintaining the condition during the optimization process. In section 6 we give the convergence result of order $O(\frac{1}{k})$, where k is the iteration counter. This convergence result is valid for all nonsmooth 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 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 the sake of notation, we apply the following convention. The

*Received by the editors January 31, 2003; accepted for publication (in revised form) January 5, 2005; published electronically September 8, 2005. 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.
http://www.siam.org/journals/siopt/16-1/42228.html

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

(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 by $\langle s, x\rangle$ 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 $A : E_1 \to E_2^*$ we define the *adjoint* operator $A^* : E_2 \to E_1^*$ 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:

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

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

$$\|Ah\|_2^* \le \|A\|_{1,2} \cdot \|h\|_1. \tag{1.1}$$

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

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

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

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

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, [4, section 2.1.3]):

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

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

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

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

for some constant $L \ge 0$. Then

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

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

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

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

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):

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

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, problem (2.1) can be written in an *adjoint* form:

$$\begin{aligned} &\max_{u \in Q_2} \phi(u), \\ &\phi(u) = -\hat{\phi}(u) + \min_{x \in Q_1} \{\langle Ax, u\rangle_2 + \hat{f}(x)\}. \end{aligned} \tag{2.3}$$

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 structures of the objects $\hat{f}$, $\hat{\phi}$, Q_1, and Q_2 are 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

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

Let μ_2 be a positive *smoothness* parameter. Consider the following function:

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

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_{μ_2} is well defined by

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

Moreover, this gradient is Lipschitz-continuous with the constant

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

(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 σ_1, and the prox-center x_0 with $d_1(x_0) = 0$. By (1.2), for any $x \in Q_1$ we have

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

Let μ_1 be a positive smoothness parameter. Consider

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

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 Danskins's theorem and Theorem 1 in [3], the gradient

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

is Lipschitz-continuous with the constant

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

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

$$\phi(u) \leq f(x), \tag{3.1}$$

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*:

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

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*

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

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). ∎

Our goal is to justify a process for updating recursively the pair $(\bar{x}, \bar{u})$, which keeps satisfying inequality (3.2) as μ_1 and μ_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*

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

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

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

∎

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*:

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

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

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

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

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

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 ∇f_{μ_2} is Lipschitz-continuous, by (1.5) we have

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

Thus, condition (3.2) can be satisfied for some primal-dual pair. Let us show how we can update points $\bar{x}$ and $\bar{u}$ in order to keep (3.2) valid for smaller values of μ_1 and μ_2. Note that in view of the symmetry of the situation, at the first step of the process we can try to decrease only μ_1, keeping μ_2 unchanged. After that, at the second step, we update μ_2 and keep μ_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 μ_1 and μ_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 μ_1^+ and μ_2, provided that τ is chosen in accordance with 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 $\hat{\phi}(\bar{u}_+) \le (1-\tau)\hat{\phi}(\bar{u}) + \tau\hat{\phi}(\hat{u})$. Therefore

$$\begin{aligned}\phi_{\mu_1^+}(\bar{u}_+) &= \min_{x\in Q_1}\left\{(1-\tau)\mu_1 d_1(x) + \langle Ax, (1-\tau)\bar{u} + \tau\hat{u}\rangle_2 + \hat{f}(x)\right\} - \hat{\phi}(\bar{u}_+)\\ &\ge \min_{x\in Q_1}\left\{(1-\tau)\left[\mu_1 d_1(x) + \langle Ax, \bar{u}\rangle_2 + \hat{f}(x) - \hat{\phi}(\bar{u})\right]_1\right.\\ &\quad \left. + \tau\left[\hat{f}(x) + \langle Ax, \hat{u}\rangle_2 - \hat{\phi}(\hat{u})\right]_2\right\}.\end{aligned}$$

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

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

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

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

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

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

Thus, putting everything together, we complete the proof as follows:

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

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

$$\langle \nabla d_1(x_0), x - x_0\rangle_1 \ge 0. \tag{5.1}$$

For x and z from Q_1 denote the *Bregman distance* between z and x as

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

If z is fixed, then $\xi(z,x)$ is strongly convex in x. Moreover, in view of (1.3)

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

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

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

As compared with the gradient mapping, the Bregman projection has several advantages. First, it is uniquely defined. Second, the optimization problem in (5.3) involves the same objects as (2.9). So, there are more chances for it to be easily solvable (see section 5.3 in [3] for an example).

Let us show that the Bregman projection can also be 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*

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

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

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

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

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 μ_1 and μ_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 μ_1 and μ_2. Let us choose $\tau \in (0,1)$ in accordance with* (4.5)

and set

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

Then the pair $(\bar{x}_+, \bar{u}_+)$ *satisfies* (3.2) *with the smoothness parameters* μ_1^+ *and* μ_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:

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

Therefore, using convexity of $\hat{f}$ and f_{μ_2}, we can estimate the term $[\,\cdot\,]_1$ as follows:

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

Thus, we can continue:

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

6. Convergence analysis. In sections 4 and 5 we have seen that the smoothness parameters μ_1 and μ_2 can be decreased by a switching strategy. Thus, in order to convert the results of Theorems 4.2 and 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:

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

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

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

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:

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

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 λ_2 is not changing, and at odd iterations λ_1 is not changing, it is convenient to put their new values in the same sequence $\{\alpha_k\}_{k=-1}^{\infty}$. Let us fix the following relations between the sequences:

$$\begin{array}{rcl} k = \quad 2l \quad : & \lambda_{1,k} = \alpha_{k-1}, & \lambda_{2,k} = \alpha_k, \\ k = \quad 2l+1 \quad : & \lambda_{1,k} = \alpha_k, & \lambda_{2,k} = \alpha_{k-1}. \end{array} \tag{6.4}$$

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

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

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

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

Also, 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:*

$$(\alpha_{k+1} - \alpha_{k-1})^2 \le \alpha_{k+1}\alpha_k\alpha_{k-1}^2, \quad k \ge 0. \tag{6.5}$$

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

$$\alpha_k = \frac{2}{k+2}, \quad k \ge -1. \tag{6.6}$$

Then

$$\tau_k = 1 - \frac{\alpha_{k+1}}{\alpha_{k-1}} = \frac{2}{k+3}, \quad k \ge 0. \tag{6.7}$$

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 with rules (6.1), (6.4), and (6.6).

METHOD 1.

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

2. Iterations ($k \geq 0$):
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* 1. *Then each pair of points satisfies the excessive gap condition. Therefore*

$$f(\bar{x}_k) - \phi(\bar{u}_k) \leq \frac{4\|A\|_{1,2}}{k+1}\sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}}. \tag{6.8}$$

Proof. In accordance with 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 Method 1. It remains to use inequality (6.2). □

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*

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

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

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

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. ∎

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

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

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

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

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

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

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 no longer need 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*

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

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:

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

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

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

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

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

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 μ_2. Let us fix $\tau \in (0,1)$ and choose $\mu_2^+ = (1-\tau)\mu_2$,*

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

Then the pair $(\bar{x}_+, \bar{u}_+)$ satisfies condition (7.4) *with smoothness parameter μ_2, provided that τ is chosen in accordance with the following relation:*

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

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

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

Since ϕ is concave, by (7.4) we obtain

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

Hence, we can finish the proof as follows:

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

□

Now we can justify the following minimization scheme.

METHOD 2.

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:

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

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 Method* 2 *satisfy the following inequality:*

$$f(\bar{x}_k) - \phi(\bar{u}_k) \leq \frac{4L_2(\phi)D_2}{(k+1)(k+2)\sigma_2}, \tag{7.10}$$

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

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

$$f(x) = \frac{1}{2}\|x\|_1^2 + \max_{1 \leq j \leq m} [f_j + \langle g_j, x - x_j \rangle_1] \quad \rightarrow \quad \min : \ x \in E_1. \tag{7.11}$$

Problems of this type arise, for example, at each iteration of the 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 Method 2.

Indeed, we can represent the objective function in (7.11) in the form (2.2) using the following objects:

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

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

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

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]):

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

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

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

Thus, Method 2 as applied to problem (7.11) converges with the following rate:

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

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

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

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

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

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

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

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

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 Method 2 as applied to problem (7.11) 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 to two.

Acknowledgment. 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*, Math. Program., 103 (2005), pp. 127–152.

[4] Yu. Nesterov, *Introductory Lectures on Convex Optimization: Basic Course*, Kluwer Academic, Boston, MA, 2004.