

Smooth minimization of non-smooth functions

Yu. Nesterov *

January 10, 2003

*CORE, Catholic University of Louvain, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium;
e-mail: nesterov@core.ucl.ac.be

Abstract

In this paper we propose a new approach for constructing efficient schemes for non-smooth convex optimization. It is based on a special smoothing technique, which can be applied to the functions with explicit max-structure. Our approach can be considered as an alternative to black-box minimization. From the viewpoint of efficiency estimates, we manage to improve the traditional bounds on the number of iterations of the gradient schemes from $O\left(\frac{1}{\epsilon^2}\right)$ to $O\left(\frac{1}{\epsilon}\right)$, keeping basically the complexity of each iteration unchanged.

Keywords: Non-smooth optimization, convex optimization, optimal methods, complexity theory, structural optimization.

1 Introduction

Motivation. Historically, the subgradient methods were the first numerical schemes for non-smooth convex minimization (see [10] and [6] for historical comments). Very soon it was proved that the efficiency estimate of these schemes is of the order

$$O\left(\frac{1}{\epsilon^2}\right), \tag{1.1}$$

where ϵ is the desired accuracy of the approximate solution (see also [3]).

Up to now some variants of these methods remain attractive for the researchers (e.g. [4, 1]). This is not too surprising since the main drawback of these schemes, the slow rate of convergence, is compensated by the very low complexity of each iteration. Moreover, it was shown in [7] that the efficiency estimate of the simplest subgradient method *cannot* be improved uniformly in dimension of space of variables. Of course, this statement is valid only for the black-box oracle model of the objective function. However, its proof is constructive. Namely, it was shown that the problem

$$\min_x \left\{ \max_{1 \leq i \leq n} x^{(i)} : \sum_{i=1}^n (x^{(i)})^2 \leq 1 \right\}$$

is difficult for all numerical schemes. This demonstration possibly explains a common belief that the worst-case complexity estimate for finding an ϵ -approximation of a minimum of piece-wise linear function by a gradient schemes is given by (1.1).

Actually, it is not the case. In practice, we never meet a pure black box model. We always know something about the structure of underlying objects. And the proper use of the structure of the problem can and does help in finding the solution.

In this paper we discuss such a possibility. Namely, we present a systematic way to approximate the initial non-smooth objective function by a function with Lipschitz-continuous gradient. After that we minimize the smooth function by an efficient gradient method of type [8], [9]. It is known that these methods have an efficiency estimate of the order $O\left(\sqrt{\frac{L}{\epsilon}}\right)$, where L is the Lipschitz constant for the gradient of the objective function. We show that in constructing a smooth ϵ -approximation of the initial function, L can be chosen of the order $\frac{1}{\epsilon}$. Thus, we end up with a gradient scheme with efficiency estimate of the order $O\left(\frac{1}{\epsilon}\right)$.

Contents. The paper is organized as follows. In Section 2 we study a simple approach for creating smooth approximations of non-smooth functions. In some aspects, our approach resembles an old technique used in the theory of Modified Lagrangians [5, 2]. It is based on the notion of an *adjoint* problem, which is a specification of the notion of a *dual* problem. An adjoint problem is not uniquely defined and its dimension is different from the dimension of the primal space. We can expect that the increase of the dimension of the adjoint space makes the structure of the adjoint problem simpler. In Section 3 we present a fast scheme for minimizing smooth convex functions. One of the advantages of this scheme consists in a possibility to use a specific norm, which is suitable for measuring the curvature of a particular objective function. This ability is similar to that one of the mirror descent methods [7, 1]. In Section 4 we apply the results of the previous section to particular problem instances: solution of a matrix game, a continuous location problem,

a variational inequality with linear operator and a problem of minimization of a piecewise linear function. For all cases we give the upper bounds on the complexity of finding ϵ -solutions for primal and dual problem. In Section 5 we discuss implementation issues and some modifications of the proposed algorithm. Preliminary computational results are given in Section 6.

Notation. In what follows we work with different primal and dual spaces equipped with corresponding norms. We use the following notation. 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 (the *dual* space) 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:

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

For an operator $A : E_1 \rightarrow E_2^*$ we define *adjoint* operator $A^* : E_2 \rightarrow E_1^*$ in the following way:

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

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

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

Clearly,

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

Hence, for any $u \in E_2$ we have

$$\|A^*u\|_1^* \leq \|A\|_{1,2} \cdot \|u\|_2. \tag{1.2}$$

2 Smooth approximations of non-differentiable functions

In this paper our main problem of interest is as follows:

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

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.

Quite often, the *structure* of the objective function in (2.1) is given explicitly. Let us assume that this structure can be described by the following *model*:

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

where the 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:

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

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

However, note that this possibility is not completely similar to (2.1) since in our case we implicitly assume that the function $\hat{\phi}(u)$ and the set Q_2 are so simple that the solution of optimization problem in (2.2) can be found in a closed form. This assumption may be not valid for the objects defining the function $\phi(u)$.

Note that for a convex function $f(x)$ the representation (2.2) is *not* uniquely defined. If we take, for example,

$$Q_2 \equiv E_2 = E_1^*, \quad \hat{\phi}(u) \equiv f_*(u) = \max_x \{\langle u, x \rangle_1 - f(x) : x \in E_1\},$$

then $\hat{f}(x) \equiv 0$, and A is equal to I , the identity operator. However, in this case the function $\hat{\phi}(u)$ may be too complicated for our goals. Intuitively, it is clear that the bigger is the dimension of space E_2 , the simpler is the structure of the adjoint objects, the function $\hat{\phi}(u)$ and the set Q_2 . Let us see that on an example.

Example 1 Consider $f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}|$. Then we can set $A = I$, $E_2 = E_1^* = R^n$ and

$$\begin{aligned} \hat{\phi}(u) &= \max_x \left\{ \langle u, x \rangle_1 - \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}| \right\} \\ &= \max_x \min_{s \in R^m} \left\{ \langle u, x \rangle_1 - \sum_{j=1}^m s^{(j)} [\langle a_j, x \rangle_1 - b^{(j)}] : \sum_{j=1}^m |s^{(j)}| \leq 1 \right\} \\ &= \min_{s \in R^m} \left\{ \sum_{j=1}^m s^{(j)} b^{(j)} : u = \sum_{j=1}^m s^{(j)} a_j, \sum_{j=1}^m |s^{(j)}| \leq 1 \right\}. \end{aligned}$$

It is clear that the structure of such a function can be very complicated.

Let us look at another possibility. Note that

$$f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}| = \max_{u \in R^m} \left\{ \sum_{j=1}^m u^{(j)} [\langle a_j, x \rangle_1 - b^{(j)}] : \sum_{j=1}^m |u^{(j)}| \leq 1 \right\}.$$

In this case $E_2 = R^m$, $\hat{\phi}(u) = \langle b, u \rangle_2$ and $Q_2 = \{u \in R^m : \sum_{j=1}^m |u^{(j)}| \leq 1\}$.

Finally, we can represent $f(x)$ also as follows:

$$f(x) = \max_{u=(u_1, u_2) \in R^{2m}} \left\{ \sum_{j=1}^m (u_1^{(j)} - u_2^{(j)}) \cdot [\langle a_j, x \rangle_1 - b^{(j)}] : \sum_{j=1}^m (u_1^{(j)} + u_2^{(j)}) = 1, u \geq 0 \right\}.$$

In this case $E_2 = R^{2m}$, $\hat{\phi}(u)$ is a linear function and Q_2 is a simplex. In Section 4.4 we will see that this representation is the best. \square

Let us show that the knowledge of the structure (2.2) can help in solving both problems (2.1) and (2.3). We are going to use this structure to construct a smooth approximation of the objective function in (2.1).

Consider a *prox-function* $d_2(u)$ of the set Q_2 . We assume that $d_2(u)$ is continuous and strongly convex on Q_2 with the convexity parameter σ_2 . Denote

$$u_0 = \arg \min_u \{d_2(u) : u \in Q_2\}.$$

Without loss of generality we assume that $d_2(u_0) = 0$. Thus, for any $u \in Q_2$ we have

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

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

$$f_\mu(x) = \max_u \{\langle Ax, u \rangle_2 - \hat{\phi}(u) - \mu d_2(u) : u \in Q_2\}. \quad (2.5)$$

Denote by $u(x)$ the optimal solution of above problem. Since function $d_2(u)$ is strongly convex, this solution is unique.

Theorem 1 *Function $f_\mu(x)$ is well defined and continuously differentiable at any $x \in E_1$. Moreover, this function is convex and its gradient*

$$\nabla f_\mu(x) = A^*u(x) \quad (2.6)$$

is Lipschitz continuous with the constant

$$L_\mu = \frac{1}{\mu\sigma_2}\|A\|_{1,2}^2.$$

Proof:

Indeed, $f_\mu(x)$ is convex as a maximum of linear functions. It is differentiable since $u(x)$ is unique. Let us prove that its gradient is Lipschitz continuous. Consider two points x_1 and x_2 . For the sake of notation, without loss of generality we assume that the functions $\hat{\phi}(\cdot)$ and $d_2(\cdot)$ are differentiable. From the first-order optimality conditions we have

$$\langle Ax_1 - \nabla \hat{\phi}(u(x_1)) - \mu \nabla d_2(u(x_1)), u(x_2) - u(x_1) \rangle_2 \leq 0,$$

$$\langle Ax_2 - \nabla \hat{\phi}(u(x_2)) - \mu \nabla d_2(u(x_2)), u(x_1) - u(x_2) \rangle_2 \leq 0.$$

Adding these inequalities and using convexity of $\hat{\phi}(\cdot)$ and strong convexity of $d_2(\cdot)$, we continue as follows:

$$\begin{aligned} & \langle A(x_1 - x_2), u(x_1) - u(x_2) \rangle_2 \\ & \geq \langle \nabla \hat{\phi}(u(x_1)) - \nabla \hat{\phi}(u(x_2)) + \mu(\nabla d_2(u(x_1)) - \nabla d_2(u(x_2))), u(x_1) - u(x_2) \rangle_2 \\ & \geq \mu \langle \nabla d_2(u(x_1)) - \nabla d_2(u(x_2)), u(x_1) - u(x_2) \rangle_2 \geq \mu\sigma_2\|u(x_1) - u(x_2)\|_2^2. \end{aligned}$$

Thus, in view of (1.2), we have

$$\begin{aligned}
(\|A^*u(x_1) - A^*u(x_2)\|_1^*)^2 &\leq \|A\|_{1,2}^2 \cdot \|u(x_1) - u(x_2)\|_2^2 \\
&\leq \frac{1}{\mu\sigma_2} \|A\|_{1,2}^2 \langle A^*(u(x_1) - u(x_2)), x_1 - x_2 \rangle_1 \\
&\leq \frac{1}{\mu\sigma_2} \|A\|_{1,2}^2 \cdot \|A^*u(x_1) - A^*u(x_2)\|_1^* \cdot \|x_1 - x_2\|_1.
\end{aligned}$$

□

Denote $D_2 = \max_u \{d_2(u) : u \in Q_2\}$ and $f_0(x) = \max_u \{\langle Ax, u \rangle_2 - \hat{\phi}(u) : u \in Q_2\}$. Then, for any $x \in E_1$ we have

$$f_\mu(x) \leq f_0(x) \leq f_\mu(x) + \mu D_2. \quad (2.7)$$

Thus, for $\mu > 0$ the function $f_\mu(x)$ can be seen as a uniform smooth approximation of the function $f_0(x)$.

In the next section we present an efficient scheme for minimizing a convex function with Lipschitz continuous gradient.

3 Optimal scheme for smooth optimization

Let us fix a function $f(x)$, which is differentiable and convex on a closed convex set $Q \subseteq E$. Assume that the gradient of this function is Lipschitz continuous:

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

(notation: $f \in C_L^{1,1}(Q)$). In this case for any $x, y \in Q$ we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2. \quad (3.1)$$

Denote by $T_Q(x) \in Q$ the optimal solution of the following minimization problem:

$$\min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2 : y \in Q \right\}. \quad (3.2)$$

If the norm $\|\cdot\|$ is not strictly convex, the problem (3.2) can have multiple solutions. In this case we stick the notation $T_Q(x)$ to any of them. In view of inequality (3.1), for any $x \in Q$ we have

$$f(T_Q(x)) \leq f(x) + \min_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2 : y \in Q \right\}. \quad (3.3)$$

Denote by $d(x)$ a *prox-function* of the set Q . We assume that $d(x)$ is continuous and strongly convex on Q with convexity parameter $\sigma > 0$. Let x_0 be the *center* of the set Q :

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

Without loss of generality assume that $d(x_0) = 0$. Thus, for any $x \in Q$ we have

$$d(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2. \quad (3.4)$$

In this section we consider an optimization scheme for solving the following problem:

$$\min_x \{f(x) : x \in Q\}, \quad (3.5)$$

with $f \in C_L^{1,1}(Q)$. For simplicity, we assume that the constant $L > 0$ is known. Recall that the standard gradient projection method at this problem converges as $O(\frac{1}{k})$, where k is the iteration counter (see, e.g. [6]).

In our scheme we update recursively two sequences of points $\{x_k\}_{k=0}^\infty \subset Q$ and $\{y_k\}_{k=0}^\infty \subset Q$ in such a way that they satisfy the following relation:

$$A_k f(y_k) \leq \psi_k \equiv \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\}, \quad (\mathcal{R}_k)$$

where $A_k = \sum_{i=0}^k \alpha_i$ and $\{\alpha_i\}_{i=0}^\infty$ are some positive step-size parameters. Let us present the way this can be done.

Indeed, for $k = 0$ let us take some $\alpha_0 \in (0, 1]$ and $y_0 = T_Q(x_0)$. Then, in view of inequalities (3.4) and (3.3), we have:

$$\begin{aligned} & \min_x \left\{ \frac{L}{\sigma} d(x) + \alpha_0 [f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle] : x \in Q \right\} \\ & \geq \alpha_0 \min_x \left\{ \frac{L}{2\alpha_0} \|x - x_0\|^2 + f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle : x \in Q \right\} \geq \alpha_0 f(y_0), \end{aligned}$$

and that is (\mathcal{R}_0) .

Denote

$$z_k = \arg \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\}.$$

Lemma 1 *Let some sequence $\{\alpha_k\}_{k=0}^\infty$ satisfy the condition:*

$$\alpha_0 \in (0, 1], \quad \alpha_{k+1}^2 \leq A_{k+1}, \quad k \geq 0. \quad (3.6)$$

Suppose that (\mathcal{R}_k) holds for some $k \geq 0$. Let us choose $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$ and

$$\begin{aligned} x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k, \\ y_{k+1} &= T_Q(x_{k+1}). \end{aligned} \quad (3.7)$$

Then the relation (\mathcal{R}_{k+1}) holds.

Proof:

Indeed, assume (\mathcal{R}_k) holds. Then, since function $d(x)$ is strictly convex, we have

$$\begin{aligned} \psi_{k+1} &= \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k+1} \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\} \\ &\geq \min_x \left\{ \psi_k + \frac{1}{2} L \|x - z_k\|^2 + \alpha_{k+1} [f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] : x \in Q \right\}. \end{aligned}$$

Further, in view of relation (\mathcal{R}_k) and the first rule in (3.7), we have

$$\begin{aligned}
& \psi_k + \alpha_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] \\
& \geq A_k f(y_k) + \alpha_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] \\
& \geq A_k [f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle] \\
& \quad + \alpha_{k+1}[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1} \rangle] \\
& = A_{k+1} f(x_{k+1}) + \alpha_{k+1} \langle \nabla f(x_{k+1}), x - z_k \rangle.
\end{aligned} \tag{3.8}$$

In view of condition (3.6), $A_{k+1}^{-1} \geq \tau_k^2$. Therefore, we can continue as follows:

$$\begin{aligned}
\psi_{k+1} & \geq A_{k+1} f(x_{k+1}) + \min_x \left\{ \frac{1}{2} L \|x - z_k\|^2 + \alpha_{k+1} \langle \nabla f(x_{k+1}), x - z_k \rangle : x \in Q \right\} \\
& = A_{k+1} \left[f(x_{k+1}) + \min_x \left\{ \frac{L}{2A_{k+1}} \|x - z_k\|^2 + \tau_k \langle \nabla f(x_{k+1}), x - z_k \rangle : x \in Q \right\} \right] \\
& \geq A_{k+1} \left[f(x_{k+1}) + \min_x \left\{ \frac{1}{2} \tau_k^2 L \|x - z_k\|^2 + \tau_k \langle \nabla f(x_{k+1}), x - z_k \rangle : x \in Q \right\} \right].
\end{aligned} \tag{3.9}$$

Finally, note that $\tau_k \in [0, 1]$. For arbitrary $x \in Q$ define

$$y = \tau_k x + (1 - \tau_k) y_k.$$

Then, in view of the first relation in (3.7) we have

$$y - x_{k+1} = \tau_k (x - z_k).$$

Hence, in view of (3.3) and the second rule in (3.7) we conclude that

$$\begin{aligned}
& \min_x \left\{ \frac{1}{2} \tau_k^2 L \|x - z_k\|^2 + \tau_k \langle \nabla f(x_{k+1}), x - z_k \rangle : x \in Q \right\} \\
& = \min_y \left\{ \frac{1}{2} L \|y - x_{k+1}\|^2 + \langle \nabla f(x_{k+1}), y - x_{k+1} \rangle : y \in \tau_k Q + (1 - \tau_k) y_k \right\} \\
& \geq \min_y \left\{ \frac{1}{2} L \|y - x_{k+1}\|^2 + \langle \nabla f(x_{k+1}), y - x_{k+1} \rangle : y \in Q \right\} \\
& \geq f(y_{k+1}) - f(x_{k+1}).
\end{aligned}$$

Combining this bound with the final estimate in (3.9) we get the result. \square

Clearly, there are many ways to satisfy the conditions (3.6). Let us give an example.

Lemma 2 For $k \geq 0$ define $\alpha_k = \frac{k+1}{2}$. Then

$$\tau_k = \frac{2}{k+3}, \quad A_k = \frac{(k+1)(k+2)}{4}, \tag{3.10}$$

and the conditions (3.6) are satisfied.

Proof:

Indeed, $\tau_k^2 = \frac{\alpha_{k+1}^2}{A_{k+1}^2} = \frac{4}{(k+3)^2} \leq \frac{4}{(k+2)(k+3)} = \frac{1}{A_{k+1}}$, and that is (3.6). \square

Now we can analyze the behavior of the following scheme.

For $k \geq 0$ **do**

1. Compute $f(x_k)$ and $\nabla f(x_k)$.

2. Find $y_k = T_Q(x_k)$.

3. Find $z_k = \arg \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \frac{i+1}{2} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\}$. (3.11)

4. Set $x_{k+1} = \frac{2}{k+3} z_k + \frac{k+1}{k+3} y_k$.

Theorem 2 *Let the sequences $\{x_k\}_{k=0}^\infty$ and $\{y_k\}_{k=0}^\infty$ be generated by the method (3.11). Then for any $k \geq 0$ we have*

$$\frac{(k+1)(k+2)}{4} f(y_k) \leq \min_x \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \frac{i+1}{2} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\}. \quad (3.12)$$

Therefore,

$$f(y_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}, \quad (3.13)$$

where x^* is an optimal solution to the problem (3.5).

Proof:

Indeed, let us choose the sequence $\{\alpha_k\}_{k=0}^\infty$ as in Lemma 2. Then, in view of Lemma 1 and convexity of $f(x)$ we have

$$A_k f(y_k) \leq \psi_k \leq \frac{L}{\sigma} d(x^*) + A_k f(x^*).$$

It remains to use (3.10). \square

Note that, in general, method (3.11) does not ensure a monotone decrease of the objective function during the minimization process. However, sometimes this property is quite useful. To achieve that, we need to introduce a minor change in the scheme.

Indeed, in the proof of Lemma 1 we need only the following condition on y_{k+1} :

$$f(y_{k+1}) \leq f(T_Q(x_{k+1})).$$

Let us change the rules of Step 2 in (3.11) as follows:

2'. Find $y'_k = T_Q(x_k)$. Compute $f(y'_k)$.

Set $y_k = \arg \min_x \{f(x) : x \in \{y_{k-1}, x_k, y'_k\}\}$.

(3.14)

Clearly, in this case we will have

$$f(y_k) \leq f(y_{k-1}) \leq \dots \leq f(x_0). \quad (3.15)$$

4 Application examples

Let us put the results of Sections 2 and 3 together. Let us assume that the function $\hat{f}(\cdot)$ in (2.2) is differentiable and its gradient is Lipschitz-continuous with some constant $M \geq 0$. Then the smoothing technique as applied to the problem (2.1) gives us the following objective function:

$$\bar{f}_\mu(x) = \hat{f}(x) + f_\mu(x) \quad \rightarrow \quad \min : x \in Q_1. \quad (4.1)$$

In view of Theorem 1, the gradient of this function is Lipschitz continuous with the constant

$$L_\mu = M + \frac{1}{\mu\sigma_2} \|A\|_{1,2}^2.$$

Let us choose some prox-function $d_1(x)$ for the set Q_1 with the convexity parameter σ_1 . For simplicity, assume that the set Q_1 is bounded:

$$\max_x \{d_1(x) : x \in Q_1\} \leq D_1.$$

Theorem 3 *Let us apply method (3.11) to the problem (4.1) with the following value of smoothness parameter:*

$$\mu = \mu(N) = \frac{2\|A\|_{1,2}}{N+1} \cdot \sqrt{\frac{D_1}{\sigma_1\sigma_2 D_2}}.$$

Then after N iterations we can generate the approximate solutions to the problems (2.1) and (2.3), namely,

$$\hat{x} = y_N \in Q_1, \quad \hat{u} = \sum_{i=0}^N \frac{2(i+1)}{(N+1)(N+2)} u(x_i) \in Q_2, \quad (4.2)$$

which satisfy the following inequality:

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\|A\|_{1,2}}{N+1} \cdot \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} + \frac{4MD_1}{\sigma_1 \cdot (N+1)^2}. \quad (4.3)$$

Thus, the complexity of finding an ϵ -solution to the problems (2.1), (2.3) by the smoothing technique does not exceed

$$4\|A\|_{1,2} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon} + \sqrt{\frac{MD_1}{\sigma_1 \epsilon}}. \quad (4.4)$$

Proof:

Let us fix an arbitrary $\mu > 0$. In view of Theorem 2, after N iterations of the method (3.11) we can deliver a point $\hat{x} = y_N$ such that

$$\bar{f}_\mu(\hat{x}) \leq \frac{L_\mu D_1}{\sigma_1(N+1)^2} + \min_x \left\{ \sum_{i=0}^N \frac{2(i+1)}{(N+1)(N+2)} [\bar{f}_\mu(x_i) + \langle \nabla \bar{f}_\mu(x_i), x - x_i \rangle_1] : x \in Q_1 \right\}. \quad (4.5)$$

Note that

$$\begin{aligned}
f_\mu(x) &= \max_u \{ \langle Ax, u \rangle_2 - \hat{\phi}(u) - \mu d_2(u) : u \in Q_2 \} \\
&= \langle Ax, u(x) \rangle_2 - \hat{\phi}(u(x)) - \mu d_2(u(x)), \\
\langle \nabla f_\mu(x), x \rangle_1 &= \langle A^* u(x), x \rangle_1.
\end{aligned}$$

Therefore

$$f_\mu(x_i) - \langle \nabla f_\mu(x_i), x_i \rangle_1 = -\hat{\phi}(u(x_i)) - \mu d_2(u(x_i)), \quad i = 0, \dots, N. \quad (4.6)$$

Thus, in view of (2.6) and (4.6) we have

$$\begin{aligned}
& \sum_{i=0}^N (i+1) [\bar{f}_\mu(x_i) + \langle \nabla \bar{f}_\mu(x_i), x - x_i \rangle_1] \\
& \leq \sum_{i=0}^N (i+1) [f_\mu(x_i) - \langle \nabla f_\mu(x_i), x_i \rangle_1] + \frac{1}{2} (N+1)(N+2) (\hat{f}(x) + \langle A^* \hat{u}, x \rangle_1) \\
& \leq - \sum_{i=0}^N (i+1) \hat{\phi}(u(x_i)) + \frac{1}{2} (N+1)(N+2) (\hat{f}(x) + \langle A^* \hat{u}, x \rangle_1) \\
& \leq \frac{1}{2} (N+1)(N+2) [-\hat{\phi}(\hat{u}) + \hat{f}(x) + \langle Ax, \hat{u} \rangle_2].
\end{aligned}$$

Hence, using (4.5), (2.3) and (2.7), we get the following bound:

$$\frac{L_\mu D_1}{\sigma_1 (N+1)^2} \geq \bar{f}_\mu(\hat{x}) - \phi(\hat{u}) \geq f(\hat{x}) - \phi(\hat{u}) - \mu D_2.$$

That is

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \mu D_2 + \frac{4 \|A\|_{1,2}^2 D_1}{\mu \sigma_1 \sigma_2 (N+1)^2} + \frac{4 M D_1}{\sigma_1 (N+1)^2}. \quad (4.7)$$

Minimizing the right-hand side of this inequality in μ we get inequality (4.3). \square

Note that the efficiency estimate (4.4) is much better than the standard bound $O(\frac{1}{\epsilon^2})$. In accordance with above theorem, for $M = 0$ the optimal dependence of the parameters μ , L_μ and N in ϵ is as follows:

$$\mu = \frac{\epsilon}{2D_2}, \quad L_\mu = \frac{D_2}{2\sigma_2} \cdot \frac{\|A\|_{1,2}^2}{\epsilon}, \quad N+1 = 4 \|A\|_{1,2} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon}. \quad (4.8)$$

Let us look now at some examples.

4.1 Minimax strategies for matrix games

Denote by Δ_n the standard simplex in R^n :

$$\Delta_n = \{ x \in R^n : x \geq 0, \sum_{i=1}^n x^{(i)} = 1 \}.$$

Let $A : R^n \rightarrow R^m$, $E_1 = R^n$ and $E_2 = R^m$. Consider the following saddle point problem:

$$\min_{x \in \Delta_n} \max_{u \in \Delta_m} \{ \langle Ax, u \rangle_2 + \langle c, x \rangle_1 + \langle b, u \rangle_2 \}. \quad (4.9)$$

From the viewpoint of players, this problem is reduced to a problem of non-smooth minimization:

$$\begin{aligned} \min_{x \in \Delta_n} f(x), \quad f(x) &= \langle c, x \rangle_1 + \max_{1 \leq j \leq m} [\langle a_j, x \rangle_1 + b^{(j)}], \\ \max_{u \in \Delta_m} \phi(u), \quad \phi(u) &= \langle b, u \rangle_2 + \min_{1 \leq i \leq n} [\langle \hat{a}_i, u \rangle_2 + c^{(i)}], \end{aligned} \quad (4.10)$$

where a_j are the rows and \hat{a}_i are the columns of the matrix A . In order to solve this pair of problems using the smoothing approach, we need to find a reasonable prox-function for the simplex. Let us compare two possibilities.

1. Euclidean distance. Let us choose

$$\begin{aligned} \|x\|_1 &= \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad d_1(x) = \frac{1}{2} \sum_{i=1}^n \left(x^{(i)} - \frac{1}{n} \right)^2, \\ \|u\|_2 &= \left[\sum_{j=1}^m (u^{(j)})^2 \right]^{1/2}, \quad d_2(x) = \frac{1}{2} \sum_{j=1}^m \left(u^{(j)} - \frac{1}{m} \right)^2. \end{aligned}$$

Then $\sigma_1 = \sigma_2 = 1$, $D_1 = 1 - \frac{1}{n} < 1$, $D_2 = 1 - \frac{1}{m} < 1$ and

$$\|A\|_{1,2} = \max_u \{ \|Ax\|_2^* : \|x\|_1 = 1 \} = \lambda_{\max}^{1/2}(A^T A).$$

Thus, in our case the estimate (4.3) for the result (4.2) can be specified as follows:

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\lambda_{\max}^{1/2}(A^T A)}{N+1}. \quad (4.11)$$

2. Entropy distance. Let us choose

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^n |x^{(i)}|, \quad d_1(x) = \ln n + \sum_{i=1}^n x^{(i)} \ln x^{(i)}, \\ \|u\|_2 &= \sum_{j=1}^m |u^{(j)}|, \quad d_2(u) = \ln m + \sum_{j=1}^m u^{(j)} \ln u^{(j)}. \end{aligned}$$

Lemma 3 *Under above choice of prox-functions we have*

$$\sigma_1 = \sigma_2 = 1, \quad D_1 = \ln n, \quad D_2 = \ln m.$$

Proof:

Note that $d_1(x)$ is two times continuously differentiable and $\langle d_1''(x)h, h \rangle = \sum_{i=1}^n \frac{(h^{(i)})^2}{x^{(i)}}$. It remains to use the following variant of Cauchy-Schwartz inequality

$$\left(\sum_{i=1}^n |h^{(i)}| \right)^2 \leq \left(\sum_{i=1}^n x^{(i)} \right) \cdot \left(\sum_{i=1}^n \frac{(h^{(i)})^2}{x^{(i)}} \right),$$

which is valid for all positive x . The reasoning for $d_2(u)$ is similar. \square

Note also, that now we get the following norm of the operator A :

$$\|A\|_{1,2} = \max_x \left\{ \max_{1 \leq j \leq m} |\langle a_j, x \rangle| : \|x\|_1 = 1 \right\} = \max_{i,j} |A^{(i,j)}|.$$

Thus, if we apply the entropy distance, the estimate (4.3) can be written as follows:

$$0 \leq f(\hat{x}) - \phi(\hat{u}) \leq \frac{4\sqrt{\ln n \ln m}}{N+1} \cdot \max_{i,j} |A^{(i,j)}|. \quad (4.12)$$

Note that typically the estimate (4.12) is much better than the Euclidean variant (4.11).

Let us write down explicitly the smooth approximation for the objective function in the first problem of (4.10) using the entropy distance. By definition,

$$\bar{f}_\mu(x) = \langle c, x \rangle_1 + \max_{u \in \Delta_m} \left\{ \sum_{j=1}^m u^{(j)} [\langle a_j, x \rangle + b^{(j)}] - \mu \sum_{j=1}^m u^{(j)} \ln u^{(j)} - \mu \ln m \right\}.$$

Let us apply the following result.

Lemma 4 *The solution of the problem*

$$\text{Find } \phi_*(s) = \max_{u \in \Delta_m} \left\{ \sum_{j=1}^m u^{(j)} s^{(j)} - \mu \sum_{j=1}^m u^{(j)} \ln u^{(j)} \right\} \quad (4.13)$$

is given by the vector $u(s) \in \Delta_m$ with the entries

$$u^{(j)}(s) = \frac{e^{s^{(j)}/\mu}}{\sum_{l=1}^m e^{s^{(l)}/\mu}}, \quad j = 1, \dots, m. \quad (4.14)$$

Therefore $\phi_*(s) = \mu \ln \left(\sum_{l=1}^m e^{s^{(l)}/\mu} \right)$.

Proof:

Indeed, the first order necessary and sufficient optimality conditions for (4.13) look as follows:

$$s^{(j)} - \mu(1 + \ln u^{(j)}) = \lambda, \quad j = 1, \dots, m,$$

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

Clearly, they are satisfied by (4.14) with $\lambda = \mu \ln \left(\sum_{l=1}^m e^{s^{(l)}/\mu} \right) - \mu$. \square

Using the result of Lemma 4, we conclude that in our case the problem (4.1) looks as follows:

$$\bar{f}_\mu(x) = \langle c, x \rangle_1 + \mu \ln \left(\frac{1}{m} \sum_{j=1}^m e^{[\langle a_j, x \rangle + b^{(j)}]/\mu} \right) \rightarrow \min : x \in \Delta_n.$$

Note that the complexity of the oracle for this problem is basically the same as that for the initial problem (4.10).

4.2 Continuous location problem

Consider the following *location* problem. There are p “cities” with “population” m_j , which are located at points $c_j \in R^n$, $j = 1, \dots, p$. We want to construct a service center at some position $x \in R^n \equiv E_1$, which minimizes the total social distance $f(x)$ to the center. On the other hand, this center must be constructed not too far from the origin.

Mathematically, the above problem can be posed as follows

$$\text{Find } f^* = \min_x \left\{ f(x) = \sum_{j=1}^p m_j \|x - c_j\|_1 : \|x\|_1 \leq \bar{r} \right\}. \quad (4.15)$$

In accordance to interpretation, it is natural to choose

$$\|x\|_1 = \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad d_1(x) = \frac{1}{2} \|x\|_1^2.$$

Then $\sigma_1 = 1$ and $D_1 = \frac{1}{2} \bar{r}^2$.

Further, the structure of the adjoint space E_2 is quite clear:

$$E_2 = (E_1^*)^p, \quad Q_2 = \{u = (u_1, \dots, u_p) \in E_2 : \|u_j\|_1^* \leq 1, j = 1, \dots, p\}.$$

Let us choose

$$\|u\|_2 = \left[\sum_{j=1}^p m_j (\|u_j\|_1^*)^2 \right]^{1/2}, \quad d_2(u) = \frac{1}{2} \|u\|_2^2.$$

Then $\sigma_2 = 1$ and $D_2 = \frac{1}{2} P$ with $P \equiv \sum_{j=1}^p m_j$. Note that the value P can be seen as the total size of the population.

It remains to compute the norm of the operator A :

$$\begin{aligned} \|A\|_{1,2} &= \max_{x,u} \left\{ \sum_{j=1}^p m_j \langle u_j, x \rangle_1 : \sum_{j=1}^p m_j (\|u_j\|_1^*)^2 = 1, \|x\|_1 = 1 \right\} \\ &= \max_{r_j} \left\{ \sum_{j=1}^p m_j r_j : \sum_{j=1}^p m_j r_j^2 = 1 \right\} = P^{1/2}. \end{aligned}$$

Putting the computed values in the estimate (4.3), we get the following rate of convergence:

$$f(\hat{x}) - f^* \leq \frac{2P\bar{r}}{N+1}. \quad (4.16)$$

Note that the value $\tilde{f}(x) = \frac{1}{P} f(x)$ corresponds to average individual expenses generated by the location x . Therefore,

$$\tilde{f}(\hat{x}) - \tilde{f}^* \leq \frac{2\bar{r}}{N+1}.$$

It is interesting that the right-hand side of this inequality is independent of any dimension. At the same time, it is clear that the reasonable accuracy for the approximate solution of the discussed problem should not be too high. Given a very low complexity of each

iteration in the scheme (3.11), the total efficiency of the proposed technique looks quite promising.

To conclude with the location problem, let us write down explicitly a smooth approximation of the objective function.

$$\begin{aligned}
f_\mu(x) &= \max_u \left\{ \sum_{j=1}^p m_j \langle u_j, x - c_j \rangle_1 - \mu d_2(u) : u \in Q_2 \right\} \\
&= \max_u \left\{ \sum_{j=1}^p m_j \left(\langle u_j, x - c_j \rangle_1 - \frac{1}{2} \mu (\|u_j\|_1^*)^2 \right) : \|u_j\|_1^* \leq 1, j = 1, \dots, p \right\} \\
&= \sum_{j=1}^p m_j \psi_\mu(\|x - c_j\|_1),
\end{aligned}$$

where the function $\psi_\mu(\tau)$, $\tau \geq 0$, is defined as follows:

$$\psi_\mu(\tau) = \max_{\gamma \in [0,1]} \{ \gamma \tau - \frac{1}{2} \mu \gamma^2 \} = \begin{cases} \frac{\tau^2}{2\mu}, & 0 \leq \tau \leq \mu, \\ \tau - \frac{\mu}{2}, & \mu \leq \tau. \end{cases} \quad (4.17)$$

4.3 Variational inequalities with linear operator

Consider a linear operator $B(w) = Bw + c: E \rightarrow E^*$, which is *monotone*:

$$\langle Bh, h \rangle \geq 0 \quad \forall h \in E_1.$$

Let Q be a bounded closed convex set in E . Then we can pose the following *variational inequality* problem:

$$\text{Find } w^* \in Q : \quad \langle B(w^*), w - w^* \rangle \geq 0 \quad \forall w \in Q. \quad (4.18)$$

Note that we can always rewrite problem (4.18) as an optimization problem. Indeed, define

$$\psi(w) = \max_v \{ \langle B(v), w - v \rangle : v \in Q \}.$$

Clearly, $\psi(w)$ is a convex function. It is well known that the problem

$$\min_w \{ \psi(w) : w \in Q \} \quad (4.19)$$

is equivalent to (4.18). For the sake of completeness let us provide this statement with a simple proof.

Lemma 5 *A point w^* is a solution to (4.19) if and only if it solves variational inequality (4.18). Moreover, for such w^* we have $\psi(w^*) = 0$.*

Proof:

Indeed, at any $w \in Q$ the function ψ is non-negative. If w^* is a solution to (4.18), then for any $v \in Q$ we have

$$\langle B(v), v - w^* \rangle \geq \langle B(w^*), v - w^* \rangle \geq 0.$$

Hence, $\psi(w^*) = 0$ and $w^* \in \text{Arg min}_{w \in Q} \psi(w)$.

Now, consider some $w^* \in Q$ with $\psi(w^*) = 0$. Then for any $v \in Q$ we have

$$\langle B(v), v - w^* \rangle \geq 0.$$

Suppose there exists some $v_1 \in Q$ such that $\langle B(w^*), v_1 - w^* \rangle < 0$. Consider the points

$$v_\alpha = w^* + \alpha(v_1 - w^*), \quad \alpha \in [0, 1].$$

Then

$$\begin{aligned} 0 &\leq \langle B(v_\alpha), v_\alpha - w^* \rangle = \alpha \langle B(v_\alpha), v_1 - w^* \rangle \\ &= \alpha \langle B(w^*), v_1 - w^* \rangle + \alpha^2 \langle B \cdot (v_1 - w^*), v_1 - w^* \rangle. \end{aligned}$$

Hence, for α small enough we get a contradiction. \square

Clearly, there are two possibilities for representing the problem (4.18), (4.19) in the form (2.1), (2.2).

1. Primal form. We take $E_1 = E_2 = E$, $Q_1 = Q_2 = Q$, $d_1(x) = d_2(x) = d(x)$, $A = B$ and

$$\hat{f}(x) = \langle b, x \rangle_1, \quad \hat{\phi}(u) = \langle b, u \rangle_1 + \langle Bu, u \rangle_1.$$

Note that the quadratic function $\hat{\phi}(u)$ is convex. For computation of function $f_\mu(x)$ we need to solve the following problem:

$$\max_u \{ \langle Bx, u \rangle_1 - \mu d(u) - \langle b, u \rangle_1 + \langle Bu, u \rangle_1 : u \in Q \}. \quad (4.20)$$

Since in our case $M = 0$, from Theorem 3 we get the following estimate for the complexity of problem (4.18):

$$\frac{4D_1 \|B\|_{1,2}}{\sigma_1 \epsilon}. \quad (4.21)$$

However, note that, because of the presence of non-trivial quadratic function in (4.20), this computation can be quite complicated. We can avoid that in the dual variant of the problem.

2. Dual form. Consider the dual variant of the problem (4.19):

$$\min_{w \in Q} \max_{v \in Q} \langle B(v), w - v \rangle = \max_{v \in Q} \min_{w \in Q} \langle B(v), w - v \rangle = - \min_{v \in Q} \max_{w \in Q} \langle B(v), v - w \rangle.$$

Thus, we can take $E_1 = E_2 = E$, $Q_1 = Q_2 = Q$, $d_1(x) = d_2(x) = d(x)$, $A = B$ and

$$\hat{f}(x) = \langle b, x \rangle_1 + \langle Bx, x \rangle_1, \quad \hat{\phi}(u) = \langle b, u \rangle_1.$$

Now the computation of function $f_\mu(x)$ becomes much simpler:

$$f_\mu(x) = \max_u \{ \langle Bx, u \rangle_1 - \mu d(u) - \langle b, u \rangle_1 : u \in Q \}.$$

It is interesting that we pay for that quite a moderate cost. Indeed, now M becomes equal to $\|B\|_{1,2}$. Hence, the complexity estimate (4.21) increases up to the following level:

$$\frac{4D_1 \|B\|_{1,2}}{\sigma_1 \epsilon} + \sqrt{\frac{D_1 \|B\|_{1,2}}{\sigma_1 \epsilon}}.$$

Note that in an important particular case of skew-symmetry of operator B , that is $B + B^* = 0$, the primal and dual variant have similar complexity.

4.4 Piece-wise linear optimization

1. Maximum of absolute values. Consider the following problem:

$$\min_x \left\{ f(x) = \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1 - b^{(j)}| : x \in Q_1 \right\}. \quad (4.22)$$

For simplicity, let us choose

$$\|x\|_1 = \left[\sum_{i=1}^n (x^{(i)})^2 \right]^{1/2}, \quad d_1(x) = \frac{1}{2} \|x\|^2.$$

Denote by A the matrix with rows a_j , $j = 1, \dots, m$. It is convenient to choose

$$E_2 = R^{2m}, \quad \|u\|_2 = \sum_{j=1}^{2m} |u^{(j)}|, \quad d_2(u) = \ln(2m) + \sum_{j=1}^{2m} u^{(j)} \ln u^{(j)}.$$

Then

$$f(x) = \max_u \{ \langle \hat{A}x, u \rangle_2 - \langle \hat{b}, u \rangle_2 : u \in \Delta_{2m} \},$$

where $\hat{A} = \begin{pmatrix} A \\ -A \end{pmatrix}$ and $\hat{b} = \begin{pmatrix} b \\ -b \end{pmatrix}$. Thus, $\sigma_1 = \sigma_2 = 1$, $D_2 = \ln(2m)$, and

$$D_1 = \frac{1}{2} \bar{r}^2, \quad \bar{r} = \max_x \{ \|x\|_1 : x \in Q_1 \}.$$

It remains to compute the norm of the operator \hat{A} :

$$\begin{aligned} \|\hat{A}\|_{1,2} &= \max_{x,u} \{ \langle \hat{A}x, u \rangle_2 : \|x\|_1 = 1, \|u\|_2 = 1 \} \\ &= \max_x \{ \max_{1 \leq j \leq m} |\langle a_j, x \rangle_1| : \|x\|_1 = 1 \} = \max_{1 \leq j \leq m} \|a_j\|_1^*. \end{aligned}$$

Putting all computed values in the estimate (4.4), we see that the problem (4.22) can be solved in

$$2\sqrt{2} \bar{r} \max_{1 \leq j \leq m} \|a_j\|_1^* \sqrt{\ln(2m)} \cdot \frac{1}{\epsilon}$$

iterations of the scheme (3.11). The standard subgradient schemes in this situation can count only on an

$$O \left(\left[\bar{r} \max_{1 \leq j \leq m} \|a_j\|_1^* \cdot \frac{1}{\epsilon} \right]^2 \right)$$

upper bound for the number of iterations.

Finally, the smooth version of the objective function in (4.22) looks as follows:

$$\bar{f}_\mu(x) = \mu \ln \left(\frac{1}{m} \sum_{j=1}^m \xi \left(\frac{1}{\mu} [\langle a_j, x \rangle + b^{(j)}] \right) \right)$$

with $\xi(\tau) = \frac{1}{2} [e^\tau + e^{-\tau}]$.

2. Sum of absolute values. Consider now the problem

$$\min_x \left\{ f(x) = \sum_{j=1}^m |\langle a_j, x \rangle_1 - b^{(j)}| : x \in Q_1 \right\}. \quad (4.23)$$

The simplest representation of function $f(x)$ looks as follows. Denote by A the matrix with the rows a_j . Let us choose

$$E_2 = R^m, \quad Q_2 = \{u \in R^m : |u^{(j)}| \leq 1, j = 1, \dots, m\},$$

$$d_2(u) = \frac{1}{2} \|u\|_2^2 = \frac{1}{2} \sum_{j=1}^m \|a_j\|_1^* \cdot (u^{(j)})^2.$$

Then the smooth version of the objective function looks as follows:

$$f_\mu(x) = \max_u \{ \langle Ax - b, u \rangle_2 - \mu d_2(u) : u \in Q_2 \} = \sum_{j=1}^m \|a_j\|_1^* \cdot \psi_\mu \left(\frac{|\langle a_j, x \rangle_1 - b^{(j)}|}{\|a_j\|_1^*} \right),$$

where the function $\psi_\mu(\tau)$ is defined by (4.17). Note that

$$\begin{aligned} \|A\|_{1,2} &= \max_{x,u} \left\{ \sum_{j=1}^m u^{(j)} \langle a_j, x \rangle_1 : \|x\|_1 \leq 1, \|u\|_2 \leq 1 \right\} \\ &\leq \max_u \left\{ \sum_{j=1}^m \|a_j\|_1^* \cdot |u^{(j)}| : \sum_{j=1}^m \|a_j\|_1^* \cdot (u^{(j)})^2 \leq 1 \right\} = D^{1/2} \equiv \left[\sum_{j=1}^m \|a_j\|_1^* \right]^{1/2}. \end{aligned}$$

On the other hand, $D_2 = \frac{1}{2}D$ and $\sigma_2 = 1$. Therefore from Theorem 3 we get the following complexity bound:

$$\frac{2}{\epsilon} \cdot \sqrt{\frac{2D_1}{\sigma_1}} \cdot \sum_{j=1}^m \|a_j\|_1^*.$$

5 Implementation issues

5.1 Computational complexity

Let us discuss the computational complexity of the method (3.11) as applied to the function $\tilde{f}_\mu(x)$. The main computations are performed at the Steps 1–3 of the algorithm.

Step 1. Call of the oracle. At this step we need to compute the solution of the following maximization problem:

$$\max_u \{ \langle Ax, u \rangle_2 - \hat{\phi}(u) - \mu d_2(u) : u \in Q_2 \}.$$

Note that from the origin of this problem we know, that this computation for $\mu = 0$ can be done in a closed form. Thus, we can expect that with properly chosen prox-function this computation is not too difficult for $\mu > 0$ also. In Section 4 we have seen three examples which confirm this belief.

Step 3. Computation of z_k . This computation consists in solving the following problem:

$$\min_x \{d_1(x) + \langle s, x \rangle_1 : x \in Q_1\}$$

for some fixed $s \in E_1^*$. If the set Q_1 and the prox-function $d_1(x)$ are simple enough, this computation can be done in a closed form (see Section 4). For some sets we need to solve an auxiliary equation with one variable. The above problem arises also in the mirror descent scheme. A discussion of different possibilities can be found in [1].

Step 2. Computation of $T_Q(x)$. Again, the complexity of this step depends on the complexity of the set Q_1 and the norm $\|\cdot\|_1$. In the literature such a computation is usually implemented with a Euclidean norm. Therefore let us discuss the general case in more detail.

Sometimes the following statement helps.

Lemma 6 *For any $g \in E^*$ and $h \in E$ we have*

$$\langle g, h \rangle_1 + \frac{1}{2}L\|h\|^2 = \max_s \left\{ \langle s, h \rangle_1 - \frac{1}{2L}(\|s - g\|^*)^2 : s \in E^* \right\}.$$

Proof:

Indeed,

$$\begin{aligned} \langle g, h \rangle_1 + \frac{1}{2}L\|h\|^2 &= \max_{r \geq 0} \{ \langle g, h \rangle_1 + r\|h\|_1 - \frac{1}{2L}r^2 \} \\ &= \max_{r, s} \{ \langle g, h \rangle_1 + \langle rs, h \rangle_1 - \frac{1}{2L}r^2 : r \geq 0, \|s\|^* = 1 \} \\ &= \max_s \{ \langle g + s, h \rangle_1 - \frac{1}{2L}(\|s\|^*)^2 : s \in E^* \} \\ &= \max_s \left\{ \langle s, h \rangle_1 - \frac{1}{2L}(\|s - g\|^*)^2 : s \in E^* \right\}. \end{aligned}$$

□

Let us check what is the complexity of computing $T_Q(x)$ in the situation discussed in Section 4.1. We need to find a solution to the problem

$$\text{Find } \psi^* = \min_x \{ \langle \bar{g}, x - \bar{x} \rangle + \frac{1}{2}L\|x - \bar{x}\|^2 : x \in \Delta_n \}, \quad (5.1)$$

where $\|x\| = \sum_{i=1}^n |x^{(i)}|$ and $\bar{x} \in \Delta_n$. Therefore, without loss of generality we can assume that

$$\min_{1 \leq i \leq n} \bar{g}^{(i)} = 0. \quad (5.2)$$

Using Lemma 6, we can rewrite the above problem as follows:

$$\begin{aligned} \psi^* &= \min_{x \in \Delta_n} \max_s \left\{ \langle s, x - \bar{x} \rangle - \frac{1}{2L}(\|s - \bar{g}\|^*)^2 \right\} \\ &= \min_{x \geq 0} \max_{s, \lambda} \left\{ \langle s, x - \bar{x} \rangle - \frac{1}{2L}(\|s - \bar{g}\|^*)^2 + \lambda(1 - \langle e_n, x \rangle) \right\} \\ &= \max_{s, \lambda} \left\{ -\langle s, \bar{x} \rangle - \frac{1}{2L}(\|s - \bar{g}\|^*)^2 + \lambda \right\} : s \geq \lambda e_n. \end{aligned}$$

Note that in our case $\|s\|^* = \max_{1 \leq i \leq n} |s^{(i)}|$. Therefore

$$-\psi^* = \min_{s, \lambda, \tau} \left\{ \langle s, \bar{x} \rangle + \frac{\tau^2}{2L} - \lambda : s^{(i)} \geq \lambda, |s^{(i)} - \bar{g}^{(i)}| \leq \tau, i = 1, \dots, n \right\}. \quad (5.3)$$

In the latter problem we can easily find the optimal values of $s^{(i)}$:

$$s_*^{(i)} = \max\{\lambda, \bar{g}^{(i)} - \tau\}, \quad i = 1, \dots, n.$$

Moreover, the feasible set of this problem is non-empty if and only if

$$\lambda \leq \bar{g}^{(i)} + \tau, \quad i = 1, \dots, n.$$

In view of (5.2), this means $\lambda \leq \tau$. Thus,

$$\begin{aligned} -\psi^* &= \min_{\tau \geq \lambda} \left\{ \sum_{i=1}^n \bar{x}^{(i)} \max\{\lambda, \bar{g}^{(i)} - \tau\} + \frac{\tau^2}{2L} - \lambda \right\} \\ &= \min_{\tau \geq \lambda} \left\{ \sum_{i=1}^n \bar{x}^{(i)} (\bar{g}^{(i)} - \tau - \lambda)_+ + \frac{\tau^2}{2L} \right\}, \end{aligned}$$

where $(a)_+ = \max\{a, 0\}$. Since the objective function of the latter problem is decreasing in λ , we conclude that $\lambda^* = \tau$.

Finally, we come to the following representation:

$$-\psi^* = \min_{\tau \geq 0} \left\{ \sum_{i=1}^n \bar{x}^{(i)} (\bar{g}^{(i)} - 2\tau)_+ + \frac{\tau^2}{2L} \right\}.$$

Clearly, its solution can be found by ordering the components of the vector $\bar{g}^{(i)}$ and checking the derivative of the objective function at the points

$$\tau_i = \frac{1}{2} \bar{g}^{(i)}, \quad i = 1, \dots, n.$$

The total complexity of this computation is of the order $O(n \ln n)$. We leave reconstruction of primal solution x^* of problem (5.1) as an exercise for the reader.

5.2 Computational stability

Our approach is based on smoothing of non-differentiable functions. In accordance to (4.8) the value of the smoothness parameter μ must be of the order of ϵ . This may cause some numerical troubles in computation of function $\tilde{f}_\mu(x)$ and its gradient. Among the examples of Section 4, only a smooth variant of objective function in Section 4.2 does not involve dangerous operations; all others need a careful implementation.

In both Section 4.1 and Section 4.4 we need a stable technique for computation of the values and the derivatives of the function

$$\eta(u) = \mu \ln \left(\sum_{j=1}^m e^{u^{(j)}/\mu} \right) \quad (5.4)$$

with very small values of the parameter μ . This can be done in the following way. Denote

$$\bar{u} = \max_{1 \leq j \leq m} u^{(j)}, \quad v^{(j)} = u^{(j)} - \bar{u}, \quad j = 1, \dots, m.$$

Then

$$\eta(u) = \bar{u} + \eta(v)$$

Note that all components of the vector v are non-negative and one of them is zero. Therefore the value $\eta(v)$ can be computed with a small numerical error. The same technique can be used for computing the gradient of this function since $\nabla \eta(u) = \nabla \eta(v)$.

The computations presented in Section 6 confirm that the proposed smoothing technique works even for a quite high accuracy.

5.3 Modified method

As we have seen, at each iteration of the method (3.11) it is necessary to solve two auxiliary minimization problems of two different types. It appears that quite often the computation of the point y_k is more complicated than that of z_k . Let us show how to modify the scheme (3.11) in order to have both auxiliary problems written in terms of prox-function $d(x)$.

For simplicity assume that $d(x)$ is differentiable. Denote by

$$\xi(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle, \quad z, x \in Q,$$

the Bregman distance between z and x . Clearly,

$$\xi(z, x) \geq \frac{1}{2} \sigma \|x - z\|^2.$$

Define the following mapping:

$$V_Q(z, g) = \arg \min_x \{ \langle g, x - z \rangle + \xi(z, x) : x \in Q \}.$$

In what follows we use the notation of Section 3.

Lemma 7 *Let sequence $\{\alpha_k\}_{k=0}^\infty$ satisfies condition (3.6). Suppose that condition (\mathcal{R}_k) holds for some $k \geq 0$. Let us choose $\gamma_k = \frac{\sigma}{L} \alpha_{k+1}$. Define*

$$\begin{aligned} x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k, \\ \hat{x}_{k+1} &= V_Q(z_k, \gamma_k \nabla f(x_{k+1})), \\ y_{k+1} &= \tau_k \hat{x}_{k+1} + (1 - \tau_k) y_k. \end{aligned} \tag{5.5}$$

Then the relation (\mathcal{R}_{k+1}) holds.

Proof:

Denote $l_k(x) \equiv \beta_k + \langle l_k, x - z_k \rangle = \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle]$. Then

$$\langle \frac{L}{\sigma} d'(z_k) + l_k, x - z_k \rangle \geq 0 \quad \forall x \in Q.$$

Hence, since $\psi_k = \frac{L}{\sigma}d(z_k) + \beta_k$, in view of inequality (3.8) we have the following:

$$\begin{aligned}
& \frac{L}{\sigma}d(x) + l_k(x) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - x_{k+1} \rangle \\
&= \frac{L}{\sigma}\xi(z_k, x) + \frac{L}{\sigma}(d(z_k) + \langle d'(z_k), x - z_k \rangle) \\
&\quad + \beta_k + \langle l_k, x - z_k \rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - x_{k+1} \rangle \\
&\geq \frac{L}{\sigma}\xi(z_k, x) + \psi_k + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - x_{k+1} \rangle \\
&\geq \frac{L}{\sigma}\xi(z_k, x) + A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - z_k \rangle.
\end{aligned}$$

Thus, using (3.6), we get the following

$$\begin{aligned}
\psi_{k+1} &\geq \min_x \left\{ \frac{L}{\sigma}\xi(z_k, x) + A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - z_k \rangle : x \in Q \right\} \\
&= \frac{L}{\sigma}\xi(z_k, \hat{x}_{k+1}) + A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), \hat{x}_{k+1} - z_k \rangle \\
&\geq \frac{1}{2}L\|\hat{x}_{k+1} - z_k\|^2 + A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), \hat{x}_{k+1} - z_k \rangle \\
&\geq A_{k+1} \left(\frac{1}{2}L\tau_k^2\|\hat{x}_{k+1} - z_k\|^2 + f(x_{k+1}) + \tau_k\langle \nabla f(x_{k+1}), \hat{x}_{k+1} - z_k \rangle \right).
\end{aligned}$$

It remains to use relation $y_{k+1} - x_{k+1} = \tau_k(\hat{x}_{k+1} - z_k)$. \square

Clearly, we can take

$$y_0 = z_0 = \arg \min_x \left\{ \frac{L}{\sigma}d(x) + \alpha_0[f(x_0) + \langle f'(x_0), x - x_0 \rangle] : x \in Q \right\}$$

for any $\alpha_0 \in (0, 1]$. In particular, we can use the sequence suggested in Lemma 2. In this case we come to the following algorithmic scheme.

1. Choose $y_0 = \arg \min_x \left\{ \frac{L}{\sigma}d(x) + \frac{1}{2}[f(x_0) + \langle f'(x_0), x - x_0 \rangle] : x \in Q \right\}$.

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

a) Find $z_k = \arg \min_x \left\{ \frac{L}{\sigma}d(x) + \sum_{i=0}^k \frac{i+1}{2}[f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] : x \in Q \right\}$.

b) Set $\tau_k = \frac{2}{k+3}$ and $x_{k+1} = \tau_k z_k + (1 - \tau_k)y_k$. (5.6)

c) Find $\hat{x}_{k+1} = V_Q(z_k, \frac{\sigma}{L}\tau_k \nabla f(x_{k+1}))$.

d) Set $y_{k+1} = \tau_k \hat{x}_{k+1} + (1 - \tau_k)y_k$.

Of course, for this method the statement of Theorem 2 holds. As an example, let us present the form of the mapping $V_Q(z, g)$ for entropy distance:

$$V_Q^{(i)}(z, g) = z^{(i)} e^{-g^{(i)}} \cdot \left[\sum_{j=1}^n z^{(j)} e^{-g^{(j)}} \right]^{-1}, \quad j = 1, \dots, n. \quad (5.7)$$

Clearly, this computation looks more attractive as compared with the strategy discussed in Section 5.1.

6 Preliminary computational results

We conclude this paper with the results of computational experiments on a random set of matrix game problems

$$\min_{x \in \Delta_n} \max_{u \in \Delta_m} \langle Ax, u \rangle_2.$$

The matrix A is generated randomly. Each of its entries is uniformly distributed in the interval $[-1, 1]$.

The goal of this numerical study is twofold. Firstly, we want to be sure that the technique discussed in this paper is stable enough to be implemented on a computer with floating point arithmetic. Secondly, it was interesting to demonstrate that the complexity of finding an ϵ -solution of the above problem indeed grows proportionally to $\frac{1}{\epsilon}$ with logarithmic factors dependent on n and m .

In order to achieve these goals we implemented the scheme (3.11) *exactly* as it is presented in the paper. We chose the parameters of the method in accordance with the recommendation (4.8). Note that for small ϵ these values become quite big. For example, if we take

$$\|A\|_{1,2} = 1, \quad n = 10^4, \quad m = 10^3, \quad \epsilon = 10^{-3},$$

then the values of parameters of the method (3.11) are as follows:

$$\mu = 0.72 \cdot 10^{-4}, \quad L_\mu = 23858.54, \quad N = 31906.$$

Thus, it was not evident that the method with such parameters could be numerically stable.

We present three sets of results. They correspond to different values of accuracy ϵ , namely to 10^{-2} , 10^{-3} and 10^{-4} . For the last value of ϵ we skip the problems of highest dimension since the general picture becomes already clear. At each step of the method we compute two matrix-vector products with matrix A . In order to check the stopping criterion, we compute the values of exact primal and dual functions at the current approximations \hat{x} and \hat{u} and check if

$$f(\hat{x}) - \phi(\hat{u}) \leq \epsilon.$$

This test is performed periodically, after one hundred (or one thousand) iterations. So, it does not increase significantly the computational time. For our computations we used a personal computer with processor Pentium 4 (2.6GHz) and frequency of RAM 1GHz. In the tables below for each problem instance we give the number of iterations, computational time in seconds and the percentage of the actual number of iterations with respect to predicted complexity N .

Looking at all three tables, we can see that the complexity of the problem indeed grows linearly with respect to $\frac{1}{\epsilon}$. Moreover, the prediction of the necessary number of iterations is very accurate. The computational time, especially for the big problems, looks quite

important. However, that is due to the fact that the matrix A is dense. In real-life problems we never meet big instances with such a level of density.

Computational results for $\epsilon = 0.01$. Table 1

$m \setminus n$	100	300	1000	3000	10000
100	808 0'', 44%	1011 0'', 49%	1112 3'', 49%	1314 12'', 54%	1415 44'', 54%
300	910 0'', 44%	1112 2'', 49%	1415 10'', 56%	1617 35'', 60%	1819 135'', 63%
1000	1112 2'', 49%	1213 8'', 48%	1415 32'', 51%	1718 115'', 58%	2020 451'', 63%

Computational results for $\epsilon = 0.001$. Table 2

$m \setminus n$	100	300	1000	3000	10000
100	6970 2'', 38%	8586 8'', 42%	9394 29'', 42%	10000 91'', 41%	10908 349'', 42%
300	7778 8'', 38%	10101 27'', 44%	12424 97'', 49%	14242 313'', 53%	15656 1162'', 54%
1000	8788 30'', 39%	11010 105'', 44%	13030 339'', 47%	15757 1083'', 53%	18282 4085'', 57%

It seems that Tables 1 and 2 present quite encouraging results. This range of accuracy is already very high for the subgradient schemes with $O(\frac{1}{2})$ complexity estimates. Of course, we can solve our problem by a cutting plane scheme, which has a linear rate of convergence. However, usually such a method decreases the gap by a constant factor in n iterations. In this aspect the results shown in the last column of Table 2 are very promising: we get three digits of accuracy after n or $2n$ iterations. At the same time, the complexity of each step in the cutting plane schemes is at least $O(\frac{1}{3}n^3)$. Therefore, even if we implement them in the smallest dimension (m), the arithmetical complexity of the computation shown in the most right-down corner of Table 3 would be equivalent to $180 \cdot 3 \cdot 2 = 1080$ iterations (since there $n = 10m$).

Computational results for $\epsilon = 0.0001$. Table 3

$m \setminus n$	100	300	1000	3000
100	67068 25'', 36%	72073 80'', 35%	74075 287'', 33%	80081 945'', 33%
300	85086 89'', 42%	92093 243'', 40%	101102 914'', 40%	112113 3302'', 41%
1000	97098 331'', 43%	100101 760'', 40%	116117 2936'', 42%	139140 11028'', 47%

The level of accuracy in Table 3 is unreachable for the standard subgradient schemes. It is quite high for cutting plane schemes also. Again, the arithmetical complexity of the process presented in the cell (3,3) of this table is equivalent to $116 \cdot 3 \cdot 2 = 696$ iterations of a cutting plane scheme in dimension $n = 1000$. That is indeed not too much for four digits of accuracy.

References

- [1] A. Ben-Tal and A. Nemirovskii. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, (SIAM, Philadelphia, 2001).
- [2] D.P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*, Academic Press, New York, 1982.
- [3] J.-L. Goffin. On the convergence rate of subgradient optimization methods. *Mathematical Programming*, 13(1977), 3, p.329-347.
- [4] J.-B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms*, vols. I and II. Springer-Verlag, 1993.
- [5] B. Polyak. On Bertsekas' method for minimization of composite function. In *Inter. Symp. Systems Opt. Analysis* (A.Bensoussan and J.L.Lions, eds.), Springer, 1979, pp. 179-186.
- [6] B. Polyak. *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York, 1987.
- [7] A. Nemirovsky and D. Yudin. *Informational Complexity and Efficient Methods for Solution of Convex Extremal Problems*, J. Wiley & Sons, New York, 1983
- [8] Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Math. Docl.), 1983, v.269, No. 3, 543-547.
- [9] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. to be published by Kluwer.
- [10] N. Shor. *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.