Operations Research Letters 38 (2010) 493–498

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

The CoMirror algorithm for solving nonsmooth constrained convex problems

Amir Beck[a,*], Aharon Ben-Tal[a], Nili Guttmann-Beck[b], Luba Tetruashvili[a]

[a] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Israel

[b] Department of Computer Science, The Academic College of Tel-Aviv Yafo, Yafo, Israel

ARTICLE INFO

Article history:
Received 17 May 2010
Accepted 26 July 2010
Available online 20 September 2010

Keywords:
Convex optimization
Gradient-based methods
Non-Euclidean projection
Mirror Descent

ABSTRACT

We introduce a first-order Mirror-Descent (MD) type algorithm for solving nondifferentiable convex problems having a combination of simple constraint set X (ball, simplex, etc.) and an additional functional constraint. The method is tuned to exploit the structure of X by employing an appropriate non-Euclidean distance-like function. Convergence results and efficiency estimates are derived. The performance of the algorithm is demonstrated by solving certain image deblurring problems.

1. Introduction

Consider the problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in X\}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, possibly nondifferentiable function satisfying a Lipschitz condition with constant L and where $X \subseteq \mathbb{R}^n$ is a compact convex set. One of the most basic methods for solving problem (1.1) is the *subgradient projection method* (see e.g., [3,13,15] and the references therein):

$$\mathbf{x}_{k+1} = P_X(\mathbf{x}_k - t_k f'(\mathbf{x}_k)).$$

Here $f'(\mathbf{x})$ denotes a subgradient of f at $\mathbf{x}$, $P_X(\cdot)$ is the Euclidean orthogonal projection operator onto the set X, and t_k is an appropriately chosen stepsize. The main advantage of the subgradient projection method is that when X is a "simple" set (e.g., ball, box, simplex or spectrahedron), then the projection operation can be executed efficiently so that the method becomes extremely simple in comparison to methods that use, for instance, second-order information. The major drawback of the subgradient projection method is its slow rate of convergence. Indeed, when X is a compact convex set, its efficiency estimate is given by (for some appropriately chosen stepsizes)

$$\min_{1\le s\le k} f(\mathbf{x}_s) - f^* \le O(1)\frac{L\,\mathrm{Diam}(X)}{\sqrt{k}}, \tag{1.2}$$

where $\mathrm{Diam}(X) = \max_{\mathbf{x},\mathbf{y}\in X} \|\mathbf{x}-\mathbf{y}\|_2$ (see [10]).

One way to improve the performance of the method is to use a non-Euclidean projection operator that reflects the specific geometry of the feasible set X. This was the idea behind the *Mirror-Descent* (MD) method originated in [10] and developed further in [2]. It was later interpreted in [1] as a subgradient non-Euclidean projection method. To understand the idea behind the MD method, let us consider the following well-known equivalent presentation of the subgradient projection method:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in X}\left\{f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{2t_k}\|\mathbf{x}-\mathbf{x}_k\|_2^2\right\},$$

that is, the next iterate is a minimizer of the linear approximation of the function at the previous iterate regularized by a prox term. A standard generalization of the subgradient projection method is devised by replacing the prox term with a distance-like function:

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in X}\left\{f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{t_k}D(\mathbf{x},\mathbf{x}_k)\right\}. \tag{1.3}$$

The term $D(\mathbf{u},\mathbf{v})$, which replaces the prox function $\frac{1}{2}\|\mathbf{u}-\mathbf{v}\|^2$, is assumed to be nonnegative and zero if and only if $\mathbf{u}=\mathbf{v}$. A popular choice for D is to take it as a Bregman distance [1,2,9,16] whose definition is now briefly recalled. Suppose we are given a strongly convex function ω on X with respect to a norm $\|\cdot\|$, which is not necessarily the Euclidean norm, and which is assumed to be continuously differentiable over some open set containing X. Let $\alpha > 0$ be the strong-convexity parameter associated with the ω and the norm, i.e.,

$$\langle\nabla\omega(\mathbf{u}) - \nabla\omega(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle \ge \alpha\|\mathbf{u}-\mathbf{v}\|^2 \quad \text{for every } \mathbf{u},\mathbf{v}\in\mathbb{R}^n.$$

The corresponding Bregman distance is:

$$B_\omega(\mathbf{u},\mathbf{v}) = \omega(\mathbf{u}) - \omega(\mathbf{v}) - \langle\nabla\omega(\mathbf{v}), \mathbf{u}-\mathbf{v}\rangle.$$

* Corresponding author. Tel.: +972 4 8294430.
E-mail addresses: becka@ie.technion.ac.il (A. Beck), abental@ie.technion.ac.il (A. Ben-Tal), becknili@mta.ac.il (N. Guttmann-Beck), lubate@tx.technion.ac.il (L. Tetruashvili).

0167-6377/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2010.08.005

By the gradient inequality we have that $B_\omega(\mathbf{u},\mathbf{v}) \geq 0$ for every $\mathbf{u},\mathbf{v} \in \mathbb{R}^n$ and that $B_\omega(\mathbf{u},\mathbf{v}) = 0$ if and only if $\mathbf{u} = \mathbf{v}$. Note that this function is neither necessarily symmetric nor does it satisfy the triangle inequality and therefore it is not a "standard" distance function. When $\omega(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|^2$ we have $B_\omega(\mathbf{u},\mathbf{v}) = \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2$, so that the Bregman distance amounts to the squared-Euclidean distance function multiplied by half. The MD method is the scheme (1.3) with $D \equiv B_\omega$ and with an "optimal" choice of the stepsizes given by $t_k = \frac{\sqrt{\Theta\alpha}}{\|f'(\mathbf{x}_k)\|_*\sqrt{k}}$ where Θ is the non-Euclidean diameter of X:

$$\Theta = \max\{B_\omega(\mathbf{u},\mathbf{v}) : \mathbf{u},\mathbf{v} \in X\}.$$

The MD method is summarized in Box I. Note that the term in the general step was rearranged and constants were ignored.

The efficiency estimate of the MD method was analyzed in [2] and is given by

$$\min_{1\leq s\leq k} f(\mathbf{x}_s) - f^* \leq O(1)\frac{L\sqrt{\Theta}}{\sqrt{\alpha}\sqrt{k}},$$

which is clearly a generalization of (1.2). We see that the efficiency estimate depends on the ratio $\frac{\Theta}{\alpha}$, so there is a freedom to choose α (a characteristic of the function ω) and Θ (a characteristic of the norm used and the shape of the feasible set X) so as to minimize $\frac{\Theta}{\alpha}$. In [1,2] it was shown that when X is a ball: $X = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq r\}$, the Euclidean setting is the most appropriate, namely $\|\cdot\| = \|\cdot\|_2$ and $\omega(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}\|^2$; when X is the unit simplex, $X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^T\mathbf{x} = 1, \mathbf{x} \geq 0\}$ (The vector $\mathbf{e}$ denotes the vector of all ones with an appropriate dimension.), the l_1 norm combined with the regularized entropy function $\omega(\mathbf{x}) = \sum_{i=1}^n (x_i+\sigma)\log(x_i+\sigma)$ is an appropriate choice (σ being a small number). We also note that the non-regularized version of the entropy can also be used. This requires a slightly more subtle analysis since the gradient of the entropy function is not defined on the boundaries of the unit simplex. Such an analysis can be found in [1].

In this paper we present an extension of the MD method for convex problems with an additional functional constraint (in addition to the simple constraint set). This method, termed the *CoMirror* method, requires at each iteration the computation of the functional constraint value and of the gradient of either the objective or constraint function (depending on the feasibility/infeasibility of the current iteration). A variation of this method in the Euclidean setting was analyzed by Nesterov in [11] and by Polyak in [12]. The roots of MD methods are in the 1983 book of Nemirovsky and Yudin [10]. The method and its convergence analysis are presented in Section 2. It is shown that in order to obtain an ε-feasible and optimal solution $O(1/\varepsilon^2)$ iterations are required. It is also shown that under additional conditions, an ε-optimal solution which is also feasible can be reached. An example from image deblurring demonstrating the potential of the method is given in Section 3.

2. Functional constraints

The Mirror-Descent method is capable of dealing with minimization problems with a simple feasible set. Our objective now is to show how the method can be adapted to tackle problems with an additional functional constraint. Suppose then that we are given a general convex problem of the form

$$(\mathrm{P}): \quad \min\{f(\mathbf{x}) : g(\mathbf{x}) \leq 0, \mathbf{x} \in X\}. \tag{2.1}$$

The following assumptions are made on problem (P):

- X is a compact convex subset of $\mathbb{R}^n$.
- f and g are convex subdifferentiable functions on X.
- The subgradients of f and g are bounded over X. Specifically, we assume that there are positive constants L_f and L_g such that $\|f'(\mathbf{x})\|_* \leq L_f$ and $\|g'(\mathbf{x})\|_* \leq L_g$ for all $\mathbf{x} \in X$ where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$:
$$\|\mathbf{y}\|_* = \sup\{\langle \mathbf{x},\mathbf{y}\rangle : \|\mathbf{x}\| \leq 1\}.$$
- The optimal set of (P) denoted by X^* is nonempty. The optimal function value is denoted by f_*.

Model (2.1) is comprised of only one convex functional constraint, but it essentially includes the case of several convex constraints. Indeed, if the feasible set is of the form $\{\mathbf{x} \in X : g_i(\mathbf{x}) \leq 0, i = 1,2,\ldots,m\}$, then it can be rewritten using a single constraint:

$$\{\mathbf{x} \in X : g(\mathbf{x}) \leq 0\},$$

by choosing $g(\mathbf{x}) = \max_{i=1,2,\ldots,m} g_i(\mathbf{x})$.

2.1. The ε-CoMirror method

We present the *ε-CoMirror* method that finds an ε-feasible and optimal solution with a similar efficiency estimate to the one devised for the MD method. The parameter ε is assumed to be nonnegative. When $\varepsilon = 0$ is zero, the "ε" prefix is omitted and the method is called the "CoMirror" method. A variation of this method when $\varepsilon > 0$ in the Euclidean setting was analyzed in [11,12]. The detailed method is given in Box II.

We will use the following notation for the set of indices of the ε-feasible solutions among the first n iterations:

$$I_n^\varepsilon = \{k \in \{1,2,\ldots,n\} : g(x_k) \leq \varepsilon\}.$$

One of the fundamental identities used in the analysis of the ε-CoMirror method (and of many other Bregman-based methods) is the following "three point lemma" which is stated via the terminology used in this paper.

Lemma 2.1 ([5]). *Let $X \subseteq \mathbb{R}^n$ be a compact convex set and let $\omega : X \to \mathbb{R}$ be continuously differentiable on some open set containing X. Then for any three points $\mathbf{a},\mathbf{b},\mathbf{c} \in X$, the following identity holds true:*

$$B_\omega(\mathbf{c},\mathbf{a}) + B_\omega(\mathbf{a},\mathbf{b}) - B_\omega(\mathbf{c},\mathbf{b}) = \langle \nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}), \mathbf{c}-\mathbf{a}\rangle. \tag{2.5}$$

Next we prove the following lemma which is the key ingredient in proving the convergence result.

Lemma 2.2. *Let $\{\mathbf{x}_k\}$ be the sequence generated by (2.2)–(2.4) (in Box II) and let $i < j$ be two integers. Then*

$$\sum_{k=i}^{j} t_k\langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{u}\rangle \leq \Theta + \frac{1}{2\alpha}\sum_{k=i}^{j} t_k^2\|\mathbf{e}_k\|_*^2 \tag{2.6}$$

for every $\mathbf{u} \in X$.

Proof. By the optimality condition for the problem on the right-hand side of (2.2) we have

$$\langle t_k\mathbf{e}_k - \nabla\omega(\mathbf{x}_k) + \nabla\omega(\mathbf{x}_{k+1}), \mathbf{u} - \mathbf{x}_{k+1}\rangle \geq 0 \quad \text{for every } \mathbf{u} \in X.$$

Hence,

$$t_k\langle \mathbf{e}_k, \mathbf{u} - \mathbf{x}_{k+1}\rangle \geq \langle \nabla\omega(\mathbf{x}_k) - \nabla\omega(\mathbf{x}_{k+1}), \mathbf{u} - \mathbf{x}_{k+1}\rangle \quad \text{for every } \mathbf{u} \in X. \tag{2.7}$$

Using (2.5) with $\mathbf{a} = \mathbf{x}_{k+1}$, $\mathbf{b} = \mathbf{x}_k$ and $\mathbf{c} = \mathbf{u}$ we have:

$$B_\omega(\mathbf{u},\mathbf{x}_{k+1}) - B_\omega(\mathbf{u},\mathbf{x}_k) + B_\omega(\mathbf{x}_{k+1},\mathbf{x}_k) = \langle \nabla\omega(\mathbf{x}_k) - \nabla\omega(\mathbf{x}_{k+1}), \mathbf{u} - \mathbf{x}_{k+1}\rangle. \tag{2.8}$$

Mirror Descent (MD)
Initialization: $\mathbf{x}_0 \in X$ arbitrary
General Step: for every $k = 0, 1, 2, \ldots$:

$$\mathbf{x}_{k+1} = \arg\min \left\{ \langle t_k f'(\mathbf{x}_k) - \nabla\omega(\mathbf{x}_k), \mathbf{x}\rangle + \omega(\mathbf{x}) : \mathbf{x} \in X \right\} \tag{1.4}$$

where

$$t_k = \frac{\sqrt{\Theta\alpha}}{\|f'(\mathbf{x}_k)\|_* \sqrt{k}}$$

Box I.

ε-**CoMirror**
Initialization: $\mathbf{x}_0 \in X$ arbitrary.
General step: for every $k = 0, 1, 2, \ldots$

$$\mathbf{x}_{k+1} = \arg\min \{ \langle t_k \mathbf{e}_k - \nabla\omega(\mathbf{x}_k), \mathbf{x}\rangle + \omega(\mathbf{x}) : \mathbf{x} \in X \}, \tag{2.2}$$

where

$$\mathbf{e}_k = \begin{cases} f'(\mathbf{x}_k) & \text{if } g(\mathbf{x}_k) \le \varepsilon, \\ g'(\mathbf{x}_k) & \text{else,} \end{cases} \tag{2.3}$$

and

$$t_k = \frac{\sqrt{\Theta\alpha}}{\|\mathbf{e}_k\|_* \sqrt{k}} \tag{2.4}$$

Box II.

Combining (2.7) and (2.8) yields

$$t_k \langle \mathbf{e}_k, \mathbf{u} - \mathbf{x}_{k+1}\rangle \ge B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) - B_\omega(\mathbf{u}, \mathbf{x}_k) + B_\omega(\mathbf{x}_{k+1}, \mathbf{x}_k).$$

Now,

$$\begin{aligned} t_k \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{u}\rangle &\le B_\omega(\mathbf{u}, \mathbf{x}_k) - B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) \\ &\quad - B_\omega(\mathbf{x}_{k+1}, \mathbf{x}_k) + t_k \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}_{k+1}\rangle \\ &\le B_\omega(\mathbf{u}, \mathbf{x}_k) - B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) - \frac{\alpha}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\ &\quad + t_k \|\mathbf{e}_k\|_* \|\mathbf{x}_k - \mathbf{x}_{k+1}\| \\ &\le B_\omega(\mathbf{u}, \mathbf{x}_k) - B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) \\ &\quad + \max_r \left\{ t_k \|\mathbf{e}_k\|_* r - \frac{\alpha}{2} r^2 \right\} \\ &= B_\omega(\mathbf{u}, \mathbf{x}_k) - B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) + \frac{1}{2\alpha} t_k^2 \|\mathbf{e}_k\|_*^2. \end{aligned}$$

Summing the above inequality for $k = i,\ i+1, \ldots, j$, we obtain

$$\sum_{k=i}^{j} t_k \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{u}\rangle \le B_\omega(\mathbf{u}, \mathbf{x}_i) - B_\omega(\mathbf{u}, \mathbf{x}_{j+1}) + \frac{1}{2\alpha} \sum_{k=i}^{j} t_k^2 \|\mathbf{e}_k\|_*^2.$$

Finally, using the inequality $B_\omega(\mathbf{u}, \mathbf{x}_i) - B_\omega(\mathbf{u}, \mathbf{x}_{j+1}) \le \Theta$, the desired result (2.6) follows. □

The efficiency estimate for the ε-CoMirror method is derived next.

Theorem 2.1. *Let $\{\mathbf{x}_k\}$ be the sequence generated by (2.2) and (2.4). Then*

$$\min \left\{ \min_{k \in I_n^\varepsilon} f(\mathbf{x}_k) - f_*, \varepsilon \right\} \le C \frac{\sqrt{\Theta} \max\{L_f, L_g\}}{\sqrt{\alpha}\sqrt{n}}, \tag{2.9}$$

where

$$C = \frac{(1 + \ln(2))}{2 - \sqrt{2}}. \tag{2.10}$$

Proof. Let $\mathbf{x}^*$ be an optimal solution of (P). Substituting $\mathbf{u} = \mathbf{x}^*$, $i = n_0$ and $j = n$ in (2.6) we obtain

$$\sum_{k=n_0}^{n} t_k \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \le \Theta + \frac{1}{2\alpha} \sum_{k=n_0}^{n} t_k^2 \|\mathbf{e}_k\|_*^2. \tag{2.11}$$

In addition,

$$\sum_{k=n_0}^{n} t_k \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \ge \left(\min_{k=n_0, n_0+1, \ldots, n} \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \right) \sum_{k=n_0}^{n} t_k. \tag{2.12}$$

If $k \in I_n^\varepsilon$, then by the (sub)gradient inequality:

$$f(\mathbf{x}^*) + \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \ge f(\mathbf{x}_k), \tag{2.13}$$

and if $k \notin I_n^\varepsilon$, then

$$\langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \ge g(\mathbf{x}^*) + \langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \ge g(\mathbf{x}_k) > \varepsilon. \tag{2.14}$$

Combining (2.11), (2.13) and (2.14), it follows that

$$\langle \mathbf{e}_k, \mathbf{x}_k - \mathbf{x}^*\rangle \ge \begin{cases} f(\mathbf{x}_k) - f_* & k \in I_n^\varepsilon, \\ \varepsilon & k \notin I_n^\varepsilon. \end{cases} \tag{2.15}$$

Combining (2.15) and (2.12) we conclude that

$$\min \left\{ \min_{k \in I_n^\varepsilon} f(\mathbf{x}_k) - f_*, \varepsilon \right\} \le \frac{\Theta + \frac{1}{2\alpha} \sum_{k=n_0}^{n} t_k^2 \|\mathbf{e}_k\|_*^2}{\sum_{k=n_0}^{n} t_k}.$$

Plugging the expression for the stepsizes, $t_k = \frac{\sqrt{\Theta\alpha}}{\|\mathbf{e}_k\|_* \sqrt{k}}$, on the right-hand side of the above inequality and using the bound $\|\mathbf{e}_k\|_* \le \max\{L_f, L_g\}$, we obtain that

$$\min \left\{ \min_{k \in I_n^\varepsilon} f(\mathbf{x}_k) - f_*, \varepsilon \right\} \le \sqrt{\frac{\Theta}{\alpha}} \max\{L_f, L_g\} \frac{1 + \frac{1}{2} \sum_{k=n_0}^{n} \frac{1}{k}}{\sum_{k=n_0}^{n} \frac{1}{\sqrt{k}}}.$$

Finally, letting $n_0 = \lfloor n/2 \rfloor$ and using the inequalities:

$$\sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \frac{1}{k} \leq 2\ln(2), \qquad \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \frac{1}{\sqrt{k}} \geq (2-\sqrt{2})\sqrt{n},$$

the efficiency estimate (2.9) is obtained. □

The result of Theorem 2.1 essentially states that in order to obtain an ε-optimal solution which is ε-feasible, $O\left(\frac{1}{\varepsilon^2}\right)$ iterations of the MD method are required. The precise statement is given in the following corollary.

Corollary 2.1. *Let $\{\mathbf{x}_k\}$ be the sequence generated by the ε-CoMirror method given by (2.2)–(2.4). Let $\beta < 1$. Then for $n \geq \left(\frac{C^2\Theta \max\{L_f, L_g\}^2}{\alpha\beta^2}\right)\frac{1}{\varepsilon^2}$ with C defined in (2.10), we have that*

$$\min_{k\in I_n^\varepsilon} f(\mathbf{x}_k) - f_* \leq \beta\varepsilon.$$

2.2. The CoMirror method ($\varepsilon = 0$)

The ε-CoMirror produces an ε-feasible and optimal solution. In some problems it is imperative to find a feasible solution, that is an $\mathbf{x} \in X$ satisfying $g(\mathbf{x}) \leq 0$ rather than an ε-feasible solution (that is, satisfying $g(\mathbf{x}) \leq \varepsilon$). Therefore, the arising question is whether the CoMirror method converges after $O(1/\varepsilon^2)$ iterations to the optimal value. The first important observation is that the CoMirror method is *exactly the same* as the ε-CoMirror method applied to

$$(\mathrm{P}_\varepsilon) \quad f_*^\varepsilon \equiv \min\{f(\mathbf{x}) : g(\mathbf{x}) \leq -\varepsilon, \mathbf{x} \in X\}.$$

Since the analysis will obviously rely on the feasibility of problems of the above form we will assume that

$$g_* \equiv \min\{g(\mathbf{x}) : \mathbf{x} \in X\} < 0, \tag{2.16}$$

so that the problem (P_ε) will be considered only for $\varepsilon < -g_*$. Let us denote the optimal function value of (P_ε) by f_ε^* and the set of feasible solutions among the first n iterations is denoted by

$$I_n \equiv I_n^0 = \{k \in \{1, 2, \ldots, n\} : g(\mathbf{x}_k) \leq 0\}.$$

We can now invoke Corollary 2.1 to obtain that for all $0 < \varepsilon < -g_*$, $\beta < 1$ and $n \geq \left(\frac{C^2\Theta \max\{L_f, L_g\}^2}{\alpha\beta^2}\right)\frac{1}{\varepsilon^2}$ it holds that

$$\min_{k\in I_n} f(\mathbf{x}_k) - f_*^\varepsilon \leq \beta\varepsilon.$$

The above result does not guarantee the $O(1/\varepsilon^2)$ complexity result since the left-hand side of the above inequality depends on f_*^ε rather than on f_*. In this section we will show that under an additional assumption of strong convexity and differentiability of the constraint function g, the difference $f_*^\varepsilon - f_*$ is upper bounded via ε and that as a consequence the $O(1/\varepsilon^2)$ result remains valid. The additional assumptions in this section are:

- There exists an optimal solution $\mathbf{x}^*$ of (P) such that $g(\mathbf{x}^*) = 0$.
- The function g is continuously differentiable on some open set containing X, and strongly convex on X (with respect to the norm $\|\cdot\|$), that is, there exists a positive constant $m_g > 0$ for which

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{m_g}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

The first assumption is rather mild. If the optimal solution is always attained at points $\tilde{\mathbf{x}}$ for which $g(\tilde{\mathbf{x}}) < 0$, then problem (P) is equivalent to the problem without the additional functional constraint, that is, equivalent to $\min\{f(\mathbf{x}) : \mathbf{x} \in X\}$, and in this case the standard MD method can be invoked. The second assumption – in particular the smoothness of g – precludes the consideration of multiple constraints for the case $\varepsilon = 0$, as g cannot in general be a maximum of functions. Note also that despite the fact that the constraint function g is smooth, problem (P) is not necessarily smooth due to the (potential) nonsmoothness of the objective function f.

The analysis of the CoMirror method relies on the analysis of perturbed problems of the form (P_ε) where $0 < \varepsilon < -g_*$. Obviously the inequality $f_* \leq f_*^\varepsilon$ holds. The next result describes an upper bound on f_*^ε in terms of f_* and is the key result in proving the convergence of the CoMirror method.

Lemma 2.3. *For every $0 < \varepsilon < -g_*$ the following inequality holds:*

$$f_*^\varepsilon \leq f_* + L_f\sqrt{\frac{2}{m_g}}\sqrt{\varepsilon}.$$

Proof. Let $\mathbf{x}^* \in X$ be an optimal solution of (P) and let $\tilde{\mathbf{x}} \in X$ be an optimal solution of (2.16). In particular we have $g(\mathbf{x}^*) = 0$. For every $t \in (0, 1)$ let us define $\mathbf{x}_t \equiv \tilde{\mathbf{x}} + t(\mathbf{x}^* - \tilde{\mathbf{x}})$. By the strong convexity of g we have

$$g(\mathbf{x}^*) - g(\mathbf{x}_t) \geq \langle \nabla g(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle + \frac{m_g}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2. \tag{2.17}$$

Since $\tilde{\mathbf{x}}$ is the optimal solution of (2.16) it follows that the function $\varphi(\eta) = g(\tilde{\mathbf{x}} + \eta(\mathbf{x}^* - \tilde{\mathbf{x}}))$ is convex and increasing over $[0, 1]$. Therefore, $\varphi'(t) = \langle \nabla g(\mathbf{x}_t), \mathbf{x}^* - \tilde{\mathbf{x}}\rangle \geq 0$, which combined with the identity $\mathbf{x}^* - \mathbf{x}_t = (1-t)(\mathbf{x}^* - \tilde{\mathbf{x}})$ and (2.17) implies that

$$g(\mathbf{x}^*) - g(\mathbf{x}_t) \geq \frac{m_g}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2. \tag{2.18}$$

By the continuity of g we obtain that there exists $t \in (0, 1)$ such that $g(\mathbf{x}_t) = -\varepsilon$, and by inequality (2.18) it follows that $\|\mathbf{x}^* - \mathbf{x}_t\| \leq \sqrt{\frac{2}{m_g}}\sqrt{\varepsilon}$. Finally,

$$\begin{aligned} f_*^\varepsilon &\leq f(\mathbf{x}_t) = f(\mathbf{x}^*) + (f(\mathbf{x}_t) - f(\mathbf{x}^*)) = f(\mathbf{x}^*) + |f(\mathbf{x}_t) - f(\mathbf{x}^*)| \\ &\leq f_* + L_f\sqrt{\frac{2}{m_g}}\sqrt{\varepsilon}. \quad \square \end{aligned}$$

The following result shows that the CoMirror method requires $O(1/\varepsilon^2)$ iterations in order to reach an ε-optimal solution.

Theorem 2.2. *Let $0 < \varepsilon < \min\left\{L_f\sqrt{-\frac{8g_*}{m_g}}, 1\right\}$. Then for every $n \geq \left(\frac{1}{\beta^2}\frac{8C^2\Theta \max\{L_f, L_g\}^2 L_f^2}{\alpha m_g}\right)\frac{1}{\varepsilon^2}$ with $\beta = \min\left\{\frac{4L_f^2}{m_g}, 0.99\right\}$, the inequality*

$$\min_{k\in I_n} f(\mathbf{x}_k) - f_* \leq \varepsilon$$

holds true.

Proof. Define $\tilde{\varepsilon} = \frac{m_g}{8L_f^2}\varepsilon^2$. Then the given upper bound on ε implies that $\tilde{\varepsilon} < -g_*$. Therefore, by Lemma 2.3 it follows that

$$f_*^{\tilde{\varepsilon}} - f_* \leq L_f\sqrt{\frac{2}{m_g}}\sqrt{\tilde{\varepsilon}} = \frac{\varepsilon}{2}.$$

Consider now the $\tilde{\varepsilon}$-CoMirror method applied on the problem $(\mathrm{P}_{\tilde{\varepsilon}})$. Then if $n \geq \frac{1}{\beta^2\tilde{\varepsilon}^2}\frac{C^2\Theta \max\{L_f, L_g\}^2}{\alpha}$ with $\beta = \min\left\{\frac{4L_f^2}{m_g}, 0.99\right\}$, it follows by Corollary 2.1 that

$$\min_{k\in I_n} f(\mathbf{x}_k) - f_*^{\tilde{\varepsilon}} \leq \beta\tilde{\varepsilon} = \beta\frac{m_g}{8L_f^2}\varepsilon^2 \leq \frac{\varepsilon^2}{2},$$

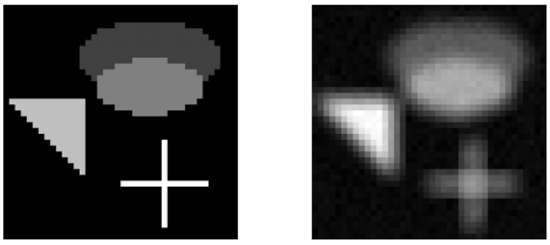

Fig. 1. Left: the original image. Right: the blurred and noisy image.

implying that

$$\min_{k\in I_n} f(\mathbf{x}_k) - f_* = \min_{k\in I_n} f(\mathbf{x}_k) - f_*^{\tilde{\varepsilon}} + f_*^{\tilde{\varepsilon}} - f_* \le \frac{\varepsilon^2}{2} + \frac{\varepsilon}{2} \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \quad \square$$

3. Application to total-variation-based deblurring

In this section we demonstrate the performance of the CoMirror method on an image deblurring problem. We consider images that are defined on rectangle domains. Let $\mathbf{b} \in \mathbb{R}^{m\times n}$ be an observed noisy and blurred image, $\mathbf{x} \in \mathbb{R}^{m\times n}$ the true (original) image to be recovered, $\mathcal{A}$ an affine map representing a blurring operator, and $\mathbf{w} \in \mathbb{R}^{m\times n}$ be a corresponding additive unknown noise satisfying the relation:

$$\mathbf{b} = \mathcal{A}(\mathbf{x}) + \mathbf{w}. \tag{3.1}$$

The problem of finding an $\mathbf{x}$ from the above relation is the basic image deblurring problem; see e.g., [8]. It is well known that the least-squares approach for this problem usually results with meaningless huge-norm solutions. Several regularization methods aimed at stabilizing the solution have been proposed in the literature. In this example we concentrate on the total-variation (TV) regularizer. There are several types of total-variation functions. One popular choice of a TV function, which will be used in our experiments, is the anisotropic TV defined by (see [4]),

$$\mathbf{x} \in \mathbb{R}^{m\times n}, \quad \mathrm{TV}(\mathbf{x}) = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\left\{|x_{i,j} - x_{i+1,j}| + |x_{i,j} - x_{i,j+1}|\right\} + \sum_{i=1}^{m-1}|x_{i,n} - x_{i+1,n}| + \sum_{j=1}^{n-1}|x_{m,j} - x_{m,j+1}|.$$

A common TV-based deblurring model consists of minimizing a total-variation function subject to a least-squares constraint; see for example [14]. Specifically, we will consider here the following optimization problem:

$$\begin{array}{ll} \min & \mathrm{TV}(\mathbf{x}) \\ \text{s.t.} & \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2 \le \rho, \\ & \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \le B, \\ & x_{ij} \ge 0, \quad i = 1,\ldots,m,\ j = 1,2,\ldots,n. \end{array} \tag{3.2}$$

In the above problem ρ is a bound on the squared norm of the noise and B is a bound on the sum of all pixels. The nonnegativity constraints on the pixels are inherent from typical image coding. Of course, problem (3.2) is a special instance of the general problem (P) (see (2.1)) with $f(\mathbf{x}) \equiv \mathrm{TV}(\mathbf{x})$, $g(\mathbf{x}) \equiv \|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|^2 - \rho$ and $X = \left\{\mathbf{x} \in \mathbb{R}^{m\times n} : \sum_{i,j} x_{ij} \le B, x_{ij} \ge 0\right\}$.

3.1. Example I

In this example we consider a small 40×40 simple test image extracted from the function blur in the "regularization toolbox" [7]. The pixels of the image were scaled to be between 0 and 1. The image goes through a 5×5 Gaussian blur with standard deviation 2 followed by an additive normally distributed noise with standard deviation 0.01 This results in the blurred and noisy image shown in Fig. 1.

We solved problem (3.2) with ρ chosen to be 10% higher than the actual squared norm of the noise and B to be also 10% higher than the actual sum of pixel values. We compared three solvers:

- *SDPT*3—The interior point method of SDPT3 [17] using the CVX interface [6].
- *MD*—The CoMirror method ($\varepsilon = 0$) with $\omega(\mathbf{x})$ chosen to be the entropy function and with the norm set to be the l_1 norm.
- *SD*—The CoMirror method with $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$. In this setting the method is almost identical to the subgradient projection method with the sole difference that when the current iterate is not feasible, the subgradient of the constraint function is used.

The results of the three methods are shown in Fig. 2. MD (after 20 000 iterations) obtained a function value of 118.73 and SDPT3 obtained a function value of 116.89. Although the value obtained by SDPT3 is slightly better, the two corresponding images look identical. SD (after 50 000 iterations) obtained a function value of 174.901 which is significantly higher than the other two methods. This difference is clearly seen by the poor reconstruction in the right image of Fig. 2.

This example demonstrates the usefulness of using non-Euclidean distances. In addition, the CPU times (on a Pentium 4, 1.8 GHz) of SDPT3 was 587 s while the CPU time of MD was 56 s. That is, SDPT3 required in this example 10 times more CPU resources, but obtained essentially the same quality of result.

3.2. Example II

The first example considered a small problem with only 1600 variables. Such a problem can still be solved by an interior point method such as SDPT3. More realistic image deblurring problems consist of hundreds of thousands or even millions of variables. In these cases, interior point methods are not a viable choice. We now consider the 512×512 "france" test image, so that the number of variables in this example is 262,144. The image goes through the same blur operator as in the first example followed by an additive normally distributed noise with standard deviation 0.01. We solved problem (3.2) with ρ and B chosen to be 10% higher than

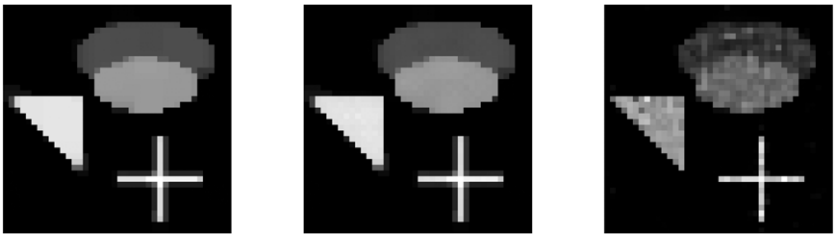

Fig. 2. Left: SDPT3 solution. Middle: MD solution. Right: SD solution.

Fig. 3. Left: blurred and noisy image. Right: MD solution.

the actual values. SDPT3 is not capable of handling such a huge problem, and the SD method was not able to even find feasible points. MD, on the other hand, was able to find a good quality solution after 423 s, which is presented in Fig. 3 together with the blurred and noisy image.

References

[1] A. Beck, M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Oper. Res. Lett. 31 (3) (2003) 167–175.
[2] A. Ben-Tal, T. Margalit, A. Nemirovski, The ordered subsets mirror descent optimization method with applications to tomography, SIAM J. Optim. 12 (1) (2001) 79–108 (electronic).
[3] D.P. Bertsekas, Nonlinear Programming, second ed., Athena Scientific, Belmont, MA, 1999.
[4] A. Chambolle, An algorithm for total variation minimization and applications, in: Mathematics and Image Analysis, J. Math. Imaging Vision 20 (1–2) (2004) 89–97 (special issue).
[5] G. Chen, M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions, SIAM J. Optim. 3 (1993) 538–543.
[6] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming (web page and software), 2009. Available at: http://stanford.edu/boyd/cvx.
[7] P.C. Hansen, Regularization tools, a Matlab package for analysis of discrete regularization problems, Numer. Algorithms 6 (1994) 1–35.
[8] P.C. Hansen, J.G. Nagy, D.P. O'Leary, Deblurring Images, in: Fundamentals of Algorithms, vol. 3, 2006.
[9] K.C. Kiwiel, Proximal minimization methods with generalized Bregman functions, SIAM J. Control Optim. 35 (1997) 1142–1168.
[10] A.S. Nemirovsky, D.B. Yudin, Problem Complexity and Method Efficiency in Optimization, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1983, Translated from the Russian and with a preface by E.R. Dawson, Wiley-Interscience Series in Discrete Mathematics.
[11] Y. Nesterov, Introductory Lectures on Convex Optimization, in: Applied Optimization, vol. 87, 2003.
[12] B.T. Polyak, A general method for solving extremal problems, Sov. Math. Dokl. 8 (1967) 593–597.
[13] B.T. Polyak, Introduction to Optimization, in: Translations Series in Mathematics and Engineering, Optimization Software Inc. Publications Division, New York, 1987, Translated from the Russian, with a foreword by Dimitri P. Bertsekas.
[14] L.I. Rudin, S.J. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268.
[15] N.Z. Shor, Minimization Methods for Nondifferentiable Functions, in: Springer Series in Computational Mathematics, vol. 3, Springer-Verlag, Berlin, 1985, Translated from the Russian by K.C. Kiwiel and A. Ruszczyński.
[16] M. Teboulle, Entropic proximal mappings with application to nonlinear programming, Math. Oper. Res. 17 (1992) 670–690.
[17] K.C. Toh, M.J. Todd, R.H. Tütüncü, SDPT3—a Matlab software package for semidefinite programming, version 1.3.