Operations Research Letters 42 (2014) 1–6

Contents lists available at ScienceDirect

Operations Research Letters

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

A fast dual proximal gradient algorithm for convex minimization and applications

Amir Beck[a,*], Marc Teboulle[b]

[a] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel
[b] School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv, Israel

ARTICLE INFO

Article history:
Received 14 July 2013
Received in revised form
25 October 2013
Accepted 25 October 2013
Available online 31 October 2013

Keywords:
Dual-based methods
Fast gradient methods
Convex optimization
Rate of convergence

ABSTRACT

We consider the convex composite problem of minimizing the sum of a strongly convex function and a general extended valued convex function. We present a dual-based proximal gradient scheme for solving this problem. We show that although the rate of convergence of the dual objective function sequence converges to the optimal value with the rate $O(1/k^2)$, the rate of convergence of the primal sequence is of the order $O(1/k)$.

1. Introduction

In this paper we focus on the nonasymptotic global rate of convergence and efficiency of a dual based proximal gradient method for minimizing the composite problem which consists of the sum of two nonsmooth convex functions, with one assumed to be strongly convex. This problem is rich enough to model many applications from diverse areas, and this will be discussed in the next section.

The literature covering both the theory and algorithms relying on the proximal technology was already vast over the last few decades and has led to fundamental algorithms, such as proximal minimization, augmented Lagrangians, splitting methods for the sum of operators, alternating direction of multipliers, and variational inequalities; see e.g., [5,11,16,18,12] for a few earlier representative works. Nowadays, the volume of research works in a wide array of new engineering applications have clearly intensified a renewed interest in proximal-based methods; see e.g., [6,9] which include several of these new applications and a comprehensive list of references.

This paper is another manifestation of the alluded current trends. Our method is a blend of old ideas combined with a very recent algorithm, demonstrating the power of Moreau proximal theory [13] when applied to optimization problems with particular structures and specific information on the problem's data. Exploiting data information, here the strong convexity of one function, we devise a novel algorithm which combines duality with the recent fast proximal gradient scheme, popularized under the name FISTA, that we recently introduced in [4]. The resulting method we obtain is called *fast dual proximal gradient* (FDPG). The idea of tackling the dual problem is not new and was developed by Tseng [20], who derived what he called *the alternating minimization method*, and which was obtained as a dual application of an algorithm introduced earlier by Gabay [11] for finding the zero of the sum of two maximal monotone operators, with one being strongly monotone. Here, by applying FISTA on the dual problem, and with essentially no extra computational cost, we derive the new method FDPG which is proven to enjoy faster global convergence rates properties than both the alternating minimization scheme as well as the classical subgradient projection algorithm when applied to the primal nonsmooth strongly convex problem, and for which we establish an improved rate of convergence over the well known $O(1/\sqrt{k})$ rate. Furthermore, as a by-product of our analysis, we can easily derive new global rate of convergence results for both the classical alternating minimization method, and the so-called dual gradient method of Uzawa [21].

Outline. Our analysis and results are developed in Sections 3 and 4, after presenting in Section 2 the optimization model we propose to study together with some interesting motivating examples. Our notations are quite standard and can be found in any convex analysis text.

[*] Corresponding author.
E-mail addresses: becka@ie.technion.ac.il (A. Beck), teboulle@post.tau.ac.il (M. Teboulle).

0167-6377/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.10.007

2. The optimization model and examples

Consider the optimization problem

$$\text{(P)} \quad \min f(\mathbf{x}) + g(\mathcal{A}\mathbf{x})$$

where $f : \mathbb{E} \to (-\infty, +\infty]$ is a proper, closed and strongly convex extended real-valued function with strong convexity parameter $\sigma > 0$ and $g : \mathbb{V} \to (-\infty, +\infty]$ is a proper, closed and convex extended real-valued function. The operator $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is a linear operator. The spaces $\mathbb{E}$, $\mathbb{V}$ are Euclidean spaces with inner products $\langle \cdot, \cdot \rangle_{\mathbb{E}}$, $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ and norms $\| \cdot \|_{\mathbb{E}}$, $\| \cdot \|_{\mathbb{V}}$. The indices will usually be omitted since the identity of the relevant space will be clear from the context. Under the properties of f and g just mentioned, problem (P) has a unique optimal solution denoted by $\mathbf{x}^*$.

Problem (P) is quite general and can model many applications from diverse areas. Following are three representatives of these applications.

Example 2.1 (*Denoising*)**.** In the denoising problem we are given a signal $\mathbf{d} \in \mathbb{E}$ which is contaminated by noise and we seek to find another vector $\mathbf{x} \in \mathbb{E}$, which on the one hand is close to $\mathbf{d}$ in the sense that the squared norm $\|\mathbf{x} - \mathbf{d}\|^2$ is small, and on the other hand, yields a small regularization term $R(\mathcal{L}\mathbf{x})$, where $\mathcal{L}$ is a linear transformation which in many applications accounts for the so-called "smoothness" of the signal and $R : \mathbb{V} \to \mathbb{R}_+$ is a given convex function that measures the magnitude of its argument. The denoising problem is then defined to be

$$\min_{\mathbf{x} \in \mathbb{E}} \|\mathbf{x} - \mathbf{d}\|^2 + \lambda R(\mathcal{L}\mathbf{x}), \tag{2.1}$$

where $\lambda > 0$ is a regularization parameter. It can be seen that problem (2.1) fits into the general model (P) by taking $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{d}\|^2$, $g(\mathbf{z}) = \lambda R(\mathbf{z})$ and $\mathcal{A} = \mathcal{L}$.

Example 2.2 (*Projection Onto the Intersection of Convex Sets*)**.** Given m closed and convex sets $C_1, C_2, \dots, C_m \subseteq \mathbb{E}$ with a nonempty intersection, and a point $\mathbf{d} \in \mathbb{E}$, the objective is to find the orthogonal projection of $\mathbf{d}$ onto the intersection of the sets, that is, the problem we consider here is

$$\min_{\mathbf{x}} \{\|\mathbf{x} - \mathbf{d}\|^2 : \mathbf{x} \in \cap_{i=1}^m C_i\}, \tag{2.2}$$

which is model (P) with $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{d}\|^2$ and $g : \mathbb{E}^m \to \mathbb{R}$ (i.e., $\mathbb{V} = \mathbb{E}^m$) defined by $g(\mathbf{z}_1, \dots, \mathbf{z}_m) = \sum_{i=1}^m \delta_{C_i}(\mathbf{z}_i)$ ($\delta_C(\cdot)$ being the indicator function of the set C). The linear operator $\mathcal{A} : \mathbb{E} \to \mathbb{E}^m$ is defined by $\mathcal{A}(\mathbf{x}) = \underbrace{(\mathbf{x}, \mathbf{x}, \dots, \mathbf{x})}_{m \text{ blocks}}$.

Example 2.3 (*Resource Allocation Problems*)**.** In many resource allocation problems we are given one-dimensional concave utility functions $u_j(x_j)$ defined over a certain interval $[m_i, M_i]$. A general model of the resource allocation problem is then

$$\begin{aligned} \max \quad & \sum_{j=1}^n u_j(x_j) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \le \mathbf{b}, \\ & x_j \in I_j \equiv [m_j, M_j], \quad j = 1, 2, \dots, n, \end{aligned} \tag{2.3}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. We will further assume that the one-dimensional functions $u_j, j = 1, 2, \dots, n$, are all strongly concave over I_j. Problem (2.3) can be cast as model (P) with $f(\mathbf{x}) = -\sum_{j=1}^n u_j(x_j)$ when $x_j \in I_j, j = 1, 2, \dots, n$, and $f(\mathbf{x}) = \infty$ otherwise, $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, and with g defined as $g(\mathbf{z}) = \delta_{(-\infty, \mathbf{b}]}(\mathbf{z})$.

3. A fast dual-based proximal gradient method

As explained in the introduction, our method is dual based and exploits the data information. We first present the dual problem and its properties. We then derive the promised algorithm in terms of the problems' data $f, g, \mathcal{A}$.

3.1. The dual problem and its properties

Problem (P) can also be written in the following constrained form:

$$\text{(P}'\text{)} \quad \min\{f(\mathbf{x}) + g(\mathbf{z}) : \mathcal{A}\mathbf{x} - \mathbf{z} = \mathbf{0}\}.$$

Associating a Lagrange dual variables vector $\mathbf{y} \in \mathbb{V}$ to the set of equality constraints in (P′), we can construct the Lagrangian of the problem

$$\begin{aligned} L(\mathbf{x}, \mathbf{z}, \mathbf{y}) &= f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} - \mathbf{z} \rangle \\ &= f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathcal{A}^T\mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle. \end{aligned} \tag{3.1}$$

Minimizing the Lagrangian with respect to $\mathbf{x}$ and $\mathbf{z}$ we obtain that the dual problem is

$$\text{(D)} \quad \max_{\mathbf{y}} \{q(\mathbf{y}) \equiv -f^*(\mathcal{A}^T\mathbf{y}) - g^*(-\mathbf{y})\}, \tag{3.2}$$

where f^* and g^* are the conjugates of f and g respectively:

$$f^*(\mathbf{y}) = \max_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\}, \qquad g^*(\mathbf{y}) = \max_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x} \rangle - g(\mathbf{x})\}.$$

We know by the strong duality theorem for convex problems (see e.g., [17]) that if there exists $\mathbf{x} \in \text{ri}(\text{dom} f)$, $\mathbf{z} \in \text{ri}(\text{dom}\, g)$ such that $\mathcal{A}\mathbf{x} = \mathbf{z}$, then strong duality holds, meaning that

$$\text{val(D)} = \text{val(P)},$$

and the optimal solution of the dual problem is attained. The strong convexity of f implies a Lipschitz gradient property of the function $f^*(\mathcal{A}^T\mathbf{x})$—a property that will be critical to our analysis. The Lipschitz constant of the gradient of $f^*(\mathbf{A}^T\mathbf{x})$ can be easily computed using a well known lemma connecting the strong convexity parameter of a convex function and the Lipschitz constant of the gradient of its conjugate [19, Proposition 12.60, p. 565].

Lemma 3.1. *The function $F(\mathbf{y}) \equiv f^*(\mathcal{A}^T\mathbf{y})$ is continuously differentiable and has a Lipschitz continuous gradient with constant $\frac{\|\mathcal{A}\|^2}{\sigma}$.*

Proof. By Proposition 12.60 from [19] it follows that f^* is continuously differentiable with a Lipschitz gradient with constant $\frac{1}{\sigma}$. Therefore, for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$:

$$\begin{aligned} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| &= \|\mathcal{A}\nabla f^*(\mathcal{A}^T\mathbf{x}) - \mathcal{A}\nabla f^*(\mathcal{A}^T\mathbf{y})\| \\ &\le \frac{1}{\sigma}\|\mathcal{A}\| \cdot \|\mathcal{A}^T\mathbf{x} - \mathcal{A}^T\mathbf{y}\| \\ &\le \frac{\|\mathcal{A}\| \cdot \|\mathcal{A}^T\|}{\sigma}\|\mathbf{x} - \mathbf{y}\| = \frac{\|\mathcal{A}\|^2}{\sigma}\|\mathbf{x} - \mathbf{y}\|. \quad \square \end{aligned}$$

We have established that the dual problem can be written as (for convenience, we consider here the equivalent minimization problem):

$$(D') \quad \min F(\mathbf{y}) + G(\mathbf{y}),$$

where

$$F(\mathbf{y}) := f^*(\mathcal{A}^T\mathbf{y}), \qquad G(\mathbf{y}) := g^*(-\mathbf{y}). \tag{3.3}$$

By Lemma 3.1 it follows that ∇F is Lipschitz continuous with constant $\frac{\|\mathcal{A}\|^2}{\sigma}$. Thus, problem (D′) consists of minimizing the sum of a smooth function F with a closed proper function G. This paves the way to apply first order proximal gradient methods on (D′) which precisely address problems of such form. This is developed in the next section where we also introduce our main scheme: a *fast* dual based proximal gradient.

3.2. The fast dual proximal gradient algorithm

We begin by recalling that the Moreau proximal map [13] of a proper closed and convex function $h : \mathbb{E} \to (-\infty, \infty]$ is defined by

$$\operatorname{prox}_h(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{E}} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|^2 \right\}.$$

Under the latter assumptions on h we also have the following well known decomposition identity (see [13, Proposition 4a, p. 280]):

$$\operatorname{prox}_h(\mathbf{z}) + \operatorname{prox}_{h^*}(\mathbf{z}) = \mathbf{z} \quad \text{for any } \mathbf{z}\in\mathbb{E}. \tag{3.4}$$

Consider the dual problem (D′). It consists of minimizing a convex objective which is the sum of a smooth function with a nonsmooth one, which is precisely the optimization model on which we can invoke the *fast* proximal gradient method, called FISTA [4], which has a faster rate of convergence guarantee.

This method, when applied to problem (D′), reads as follows:

- *Initialization:* $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$, $\mathbf{w}_1 = \mathbf{y}_0 \in \mathbb{V}$, $t_1 = 1$.
- *General step* ($k \geq 1$):

$$\mathbf{y}_k = \operatorname{prox}_{\frac{1}{L}G}\left(\mathbf{w}_k - \frac{1}{L}\nabla F(\mathbf{w}_k)\right) \tag{3.5}$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \tag{3.6}$$

$$\mathbf{w}_{k+1} = \mathbf{y}_k + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{y}_k - \mathbf{y}_{k-1}). \tag{3.7}$$

Note that with the choice $t_k = 1$ for all k, this method reduces to the original proximal gradient scheme, but which is known to be significantly slower; see [4].

The rate of convergence of $O(1/k^2)$ for the dual objective function $q(\cdot)$ given in (D) is now recalled.

Theorem 3.1 *([4, Theorem 4.4]).* *Let $\{\mathbf{y}_k\}$ be the sequence generated by (3.5)–(3.7) with $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$ and $\mathbf{w}_1 = \mathbf{y}_0 \in \mathbb{V}$, $t_1 = 1$, and let $\mathbf{y}^*$ be any optimal dual solution of problem* (D). *Then for any $k \geq 1$,*

$$q(\mathbf{y}^*) - q(\mathbf{y}_k) \leq \frac{2L\|\mathbf{y}_0 - \mathbf{y}^*\|^2}{k^2}.$$

Our objective is now to rewrite the iterations (3.5) in terms of the data of the problem $(f, g, \mathcal{A})$ which will lead to the fast dual proximal gradient method for solving (P).

Lemma 3.2. *The iteration given in* (3.5) *by* $\mathbf{y}_k = \operatorname{prox}_{\frac{1}{L}G}\left(\mathbf{w}_k - \frac{1}{L}\nabla F(\mathbf{w}_k)\right)$ *is equivalent to* $\mathbf{y}_k = \mathbf{w}_k - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k)$, *with*

$$\mathbf{u}_k = \operatorname*{argmax}_{\mathbf{x}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{w}_k\rangle - f(\mathbf{x}) \right\}, \tag{3.8}$$

$$\mathbf{v}_k = \operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{w}_k). \tag{3.9}$$

Proof. By the definition of F (Eq. (3.3)), it follows that

$$\nabla F(\mathbf{w}_k) = \mathcal{A}\nabla f^*(\mathcal{A}^T\mathbf{w}_k). \tag{3.10}$$

Since f is strongly convex, its conjugate is continuously differentiable, and hence $\mathbf{u} \in \partial f(\mathbf{v})$ if and only if $\mathbf{v} = \nabla f^*(\mathbf{u})$ (see e.g., [17, Corollary 23.5.1]). As a consequence, we thus obtain

$$\nabla f^*(\mathcal{A}^T\mathbf{w}_k) = \operatorname*{argmax}_{\mathbf{x}\in\mathbb{E}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{w}_k\rangle - f(\mathbf{x}) \right\}. \tag{3.11}$$

Let $\mathbf{d}_k := \mathbf{w}_k - \frac{1}{L}\nabla F(\mathbf{w}_k)$. Then by (3.10) and (3.11) the computation of $\mathbf{d}_k$ can be written as

$$\mathbf{u}_k = \operatorname*{argmax}_{\mathbf{x}\in\mathbb{E}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{w}_k\rangle - f(\mathbf{x}) \right\}, \tag{3.12}$$

$$\mathbf{d}_k = \mathbf{w}_k - \frac{1}{L}\mathcal{A}\mathbf{u}_k \tag{3.13}$$

and the iteration (3.5) reads as $\mathbf{y}_k = \operatorname{prox}_{\frac{1}{L}G}(\mathbf{d}_k)$. Then invoking the identity (3.4) with $h(\mathbf{y}) := \frac{1}{L}G(\mathbf{y})$ we obtain

$$\mathbf{y}_k = \mathbf{d}_k - \operatorname{prox}_{h^*}(\mathbf{d}_k). \tag{3.14}$$

To complete the proof, we need to compute $\operatorname{prox}_{h^*}$. Now, by (3.3) we have $h(\mathbf{y}) = \frac{1}{L}g^*(-\mathbf{y})$. Using the definition of the conjugate (recalling that here $g = g^{**}$), and of the proximal map, an easy computation shows that $h^*(\mathbf{v}) = 1/Lg(-L\mathbf{v})$ for any $\mathbf{v} \in \mathbb{E}$ and that

$$\operatorname{prox}_{h^*}(\mathbf{d}) = -\frac{1}{L}\operatorname{prox}_{Lg}(-L\mathbf{d}) \quad \text{for any } \mathbf{d} \in \mathbb{V}. \tag{3.15}$$

Therefore, using (3.15) in (3.14) we obtain

$$\begin{aligned}
\mathbf{y}_k &= \mathbf{d}_k + \frac{1}{L}\operatorname{prox}_{Lg}(-L\mathbf{d}_k) \\
&= \mathbf{w}_k - \frac{1}{L}\mathcal{A}\mathbf{u}_k + \frac{1}{L}\operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{w}_k) \quad \text{using (3.13)} \\
&= \mathbf{w}_k - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k),
\end{aligned}$$

with $\mathbf{v}_k = \operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{w}_k)$. □

Thanks to Lemma 3.2, we are ready to rewrite the method in terms of the data of the problem, meaning $(f, g, \mathcal{A})$.

The Fast Dual-Based Proximal Gradient Method (FDPG)

Input: $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$ - an upper bound on the Lipschitz constant of ∇F

Step 0. Take $\mathbf{w}_1 = \mathbf{y}_0 \in \mathbb{V}$, $t_1 = 1$.

Step k. ($k \geq 0$) Compute

$$\mathbf{u}_k = \operatorname*{argmax}_{\mathbf{x}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{w}_k\rangle - f(\mathbf{x}) \right\} \tag{3.16}$$

$$\mathbf{v}_k = \operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{w}_k) \tag{3.17}$$

$$\mathbf{y}_k = \mathbf{w}_k - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k).$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$$

$$\mathbf{w}_{k+1} = \mathbf{y}_k + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{y}_k - \mathbf{y}_{k-1}).$$

As noted earlier, in the special case $t_k \equiv 1$ for all k, the method corresponds to the usual proximal gradient when applied to the dual (D′), and in that case the three last steps of FDPG collapse to (after performing an index shift) $\mathbf{y}_k = \mathbf{y}_{k-1} - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k)$, $\mathbf{w}_{k+1} = \mathbf{y}_k$, and the resulting main steps of the algorithm read as follows:

$$\mathbf{u}_k = \operatorname*{argmax}_{\mathbf{x}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{y}_k\rangle - f(\mathbf{x}) \right\} \tag{3.18}$$

$$\mathbf{v}_k = \operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{y}_k) \tag{3.19}$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k). \tag{3.20}$$

This recovers the alternating minimization algorithm of Tseng [20] which was obtained as a dual application of an algorithm introduced earlier by Gabay [11] for finding the zero of the sum of two maximal monotone operators, with one being strongly monotone. Thus, through the FDPG method we obtain a natural fast version of the alternating minimization scheme which not only allows us to derive the global convergence rate results for the classical alternating minimization, but more importantly will be shown to enjoy significantly better convergence rate properties. We note that one of the main advantages of the alternating minimization scheme, and hence of the fast version, is that it can beneficially exploit the separability of a given function f, thanks to the maximization step (3.16) which can decompose accordingly. This is in sharp contrast with augmented Lagrangian-based schemes and related Alternating Direction of Multipliers whereby the presence of a coupling quadratic term prevents to exploit such a refined decomposition for given separable function; see e.g., [12,6].

Remark 3.1. Note that in the special case $g(\mathbf{z}) = \delta_{\{\mathbf{b}\}}(\mathbf{z})$, the problem (P) reduces to minimizing a strongly convex function f over linear constraints $\mathcal{A}\mathbf{x} = \mathbf{b}$ (similarly for inequality), and thus the FDPG yields naturally a fast version of the so-called Uzawa method [21].

Remark 3.2. The non-accelerated method was employed in the context of total variation-based image denoising in [7], and the corresponding realization of the FDPG method was considered in [3].

Remark 3.3. A different non-accelerated dual proximal-based method was considered in [8], and where the convergence of the sequence was derived. The problem studied in [8] is essentially the same besides the fact the strongly convex function f was given as a sum of convex function and a squared Euclidean term.

Remark 3.4. The algorithm FDGP assumes that the strong convexity parameter σ is known or can be well approximated. If σ is unknown, it is still possible to apply the algorithm and preserves its convergence properties by using a sort of backtracking procedure which is very similar to the one described in [4], and thus we omit the details.

4. Rate of convergence analysis

In this section we establish two different types of global rate of convergence results. First, we consider a primal sequence generated by the fast dual gradient proximal algorithm FDPG and we prove that this sequence converges at the rate $O(1/k)$. We then compare our algorithm versus the subgradient projection method which is direct scheme applied to the original primal formulation of the problem (P). We show that even when taking into account the strong convexity of the objective function, our method achieves the superior rate of convergence both in function values and in the sequences.

4.1. Rate of convergence of the primal sequence

Let $\{\mathbf{y}_k\}$ be the sequence generated by the fast dual proximal gradient method. Then we know by Theorem 3.1 that $q(\mathbf{y}_k)$ converges to $q(\mathbf{y}^*)$ in a rate of $O(1/k^2)$. Given as input a dual sequence $\{\mathbf{y}_k\}$ generated by FDPG, a primal sequence can be defined naturally as

$$\mathbf{x}_k = \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \mathcal{A}^T\mathbf{y}_k \rangle - f(\mathbf{x}) \right\}. \tag{4.1}$$

The sequence $\{\mathbf{x}_k\}$ is contained in $\operatorname{dom}(f)$, but is not necessarily feasible since $\mathcal{A}\mathbf{x}_k$ might not belong to $\operatorname{dom} g$. This infeasibility is a common property of dual-based methods. We will now establish a rate of convergence of the primal sequence $\{\mathbf{x}_k\}$ to the optimal solution $\mathbf{x}^*$.

Theorem 4.1. *Let $\{\mathbf{y}_k\}$ be the sequence generated by the fast dual proximal gradient method and let $\{\mathbf{x}_k\}$ be the corresponding primal sequence defined by (4.1). Then*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le 2\sqrt{\frac{L}{\sigma}}\frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k}. \tag{4.2}$$

Proof. Let $\mathbf{y_k}$ be the output of FDPG at iteration k. Let us define an additional sequence $\{\mathbf{z}_k\}$ given by

$$\mathbf{z}_k \in \underset{\mathbf{z}\in\mathbb{V}}{\operatorname{argmin}}\{\langle \mathbf{y}_k, \mathbf{z}\rangle + g(\mathbf{z})\}. \tag{4.3}$$

Let $k \ge 1$. Define

$$h_1(\mathbf{x}) = f(\mathbf{x}) - \langle \mathcal{A}^T\mathbf{y}_k, \mathbf{x}\rangle, \qquad h_2(\mathbf{z}) = g(\mathbf{z}) + \langle \mathbf{y}_k, \mathbf{z}\rangle.$$

Then by (3.1) it follows that

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}_k) = h_1(\mathbf{x}) + h_2(\mathbf{z}) \quad \text{for all } \mathbf{x} \in \mathbb{E}, \mathbf{z} \in \mathbb{V}. \tag{4.4}$$

By (4.1) and (4.3) it follows that

$$\mathbf{x}_k = \underset{\mathbf{x}\in\mathbb{E}}{\operatorname{argmin}}\, h_1(\mathbf{x}), \tag{4.5}$$

$$\mathbf{z}_k \in \underset{\mathbf{z}\in\mathbb{V}}{\operatorname{argmin}}\, h_2(\mathbf{z}). \tag{4.6}$$

By the strong convexity of f, it follows that the function $h_1(\mathbf{x})$ is strongly convex with parameter $\sigma > 0$. Therefore, by (4.5), we have that

$$h_1(\mathbf{x}) - h_1(\mathbf{x}_k) \ge \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2, \quad \text{for all } \mathbf{x} \in \mathbb{E}. \tag{4.7}$$

By (4.6) we have for any $z \in \mathbb{V}$,

$$h_2(\mathbf{z}) - h_2(\mathbf{z}_k) \ge 0. \tag{4.8}$$

Summing inequalities (4.7), (4.8) and (4.4) we obtain

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) \ge \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2, \quad \text{for all } \mathbf{x} \in \mathbb{E}, \mathbf{z} \in \mathbb{V}.$$

In particular, substituting $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{z}^* = \mathcal{A}\mathbf{x}^*$ we have

$$L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) \ge \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2. \tag{4.9}$$

Now, by from (4.1) and (4.3) and the definition of q given in (3.2) we obtain

$$\begin{aligned} L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) &= q(\mathbf{y}_k), \\ L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}_k) &= f(\mathbf{x}^*) + g(\mathbf{z}^*) - \langle \mathbf{y}_k, \mathcal{A}\mathbf{x}^* - \mathbf{z}^*\rangle \\ &= f(\mathbf{x}^*) + g(\mathcal{A}\mathbf{x}^*) = q(\mathbf{y}^*), \end{aligned}$$

where the last equality follows from strong duality. Therefore, from (4.9) and Theorem 3.1 we get

$$\frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \le q(\mathbf{y}^*) - q(\mathbf{y}_k) \le \frac{2L\|\mathbf{y}_0 - \mathbf{y}^*\|^2}{k^2},$$

and hence

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le 2\sqrt{\frac{L}{\sigma}}\frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k}. \quad \Box$$

If L is chosen to be exactly $\frac{\|\mathcal{A}\|^2}{\sigma}$, then result (4.2) will read as

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le 2\frac{\|\mathcal{A}\|}{\sigma}\frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k}.$$

Remark 4.1. A similar rate of convergence for a different dual-based method that uses Nesterov's fast gradient method from [15] was derived in [22,10] for the special case of total variation image processing problems. The analysis techniques developed in these works are rather intricate and completely different from ours, and the resulting new fast version of the alternating minimization FDPG we propose here was not derived and analyzed, since here we rely on FISTA which is simpler than the more computationally demanding scheme of [15] which involves an accumulated history of the past iterates and two prox operations per iteration.

As a byproduct of this analysis, for the special case $t_k \equiv 1$, we can immediately derive the global rate of convergence of the alternating minimization algorithm for the sequence $\mathbf{x}_k$, a result which does not seem to have been established in the literature, showing that this method is slower than FDPG by an order of magnitude. With the same proof as the one of Theorem 4.1, but now invoking instead the rate of convergence for the usual proximal gradient (c.f. [4, Theorem 3.1, p. 10], instead of Theorem 3.1) we obtain the following corollary.

Corollary 4.1. *Let $\{\mathbf{y}_k\}$ be the sequence generated by the alternating minimization method and let $\{\mathbf{x}_k\}$ be the corresponding primal sequence defined by (4.1). Then*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le \sqrt{\frac{L}{2\sigma}}\frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{\sqrt{k}}.$$

4.2. Comparison to subgradient projection

Defining $H(\mathbf{x}) := f(\mathbf{x}) + g(\mathcal{A}\mathbf{x})$, we can rewrite problem (P) as

$$\min\{H(\mathbf{x}) : \mathbf{x} \in X\}, \tag{4.10}$$

where the closed and convex feasibility set X is given by

$$X = \{\mathbf{x} \in \mathbb{E} : \mathbf{x} \in \mathrm{dom}(f), \mathcal{A}\mathbf{x} \in \mathrm{dom}(g)\}.$$

The optimal value of problem (4.10) is denoted by $H^* = H(\mathbf{x}^*)$, where $\mathbf{x}^* \in X$ is the optimal solution of (4.10). An alternative approach for solving this primal formulation (4.10) of (P) is to use a subgradient projection method. For that we will make the usual assumptions needed to analyze the rate of convergence of the subgradient projection algorithm, namely that

- H is subdifferentiable over X.
- $\gamma := \max_{\mathbf{x}\in X} \max_{\mathbf{d}\in\partial H(\mathbf{x})} \|\mathbf{d}\| < \infty$.

The subgradient projection method can be written as

$$\mathbf{x}_{k+1} = P_X(\mathbf{x}_k - t_k H'(\mathbf{x}_k)), \quad H'(\mathbf{x}_k) \in \partial H(\mathbf{x}_k),\ k \ge 0 \tag{4.11}$$

where $t_k > 0$ are appropriately chosen stepsizes.

Before comparing the two approaches, we note that the subgradient projection method requires the ability to compute the orthogonal projection onto the feasible set X, which in some cases, such as the one described in Example 2.2, is as difficult as the solution of (P) itself. In addition, the convergence results of the method rely on an additional assumption that the feasible set X is bounded.

By choosing the stepsizes in an appropriate way, it can be shown that the sequence $H_{(n)} \equiv \min\{H(\mathbf{x}_k) : k = 1, \ldots, n\}$ converges to the optimal value H^* at a rate of $O(1/\sqrt{n})$; see e.g., [14,2,1]. The $O(1/\sqrt{n})$ rate is clearly worse than the $O(1/n^2)$ rate of convergence established for the dual function values of the fast dual proximal gradient method, but in a sense this comparison is not fair for two reasons. First, the FDPG is a dual method tackling the constrained equivalent reformulation of (P) and not the direct primal formulation (4.10); second, the subgradient projection method does not exploit the strong convexity of the objective function H. In the next theorem we show how the rate of convergence of the subgradient projection method can be improved when the stepsizes are chosen in a specific way, and where the strong convexity is exploited in the analysis.

Theorem 4.2. *Let $\{\mathbf{x}_k\}$ be the sequence generated by the subgradient projection method with $\mathbf{x}_0 \in X$ and $t_k = \frac{1}{k\sigma}$. Then for all $n \ge 1$*

$$\frac{\sum_{k=1}^n H(\mathbf{x}_k)}{n} - H^* \le \frac{\gamma^2 \ln(n+1)}{2\sigma n}.$$

In addition, for all $n \ge 2$ the inequality

$$\|\mathbf{x}_n - \mathbf{x}^*\| \le \frac{\gamma}{\sqrt{\sigma}}\sqrt{\frac{\ln(n)}{n-1}} \tag{4.12}$$

holds true.

Proof. Let $\mathbf{x}^*$ be the optimal solution of problem (P). Then

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|P_X(\mathbf{x}_k - t_k H'(\mathbf{x}_k)) - P_X(\mathbf{x}^*)\|^2\\ &\le \|\mathbf{x}_k - t_k H'(\mathbf{x}_k) - \mathbf{x}^*\|^2\\ &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2t_k\langle H'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + t_k^2\|H'(\mathbf{x}_k)\|^2,\end{aligned}$$

where the inequality follows from the nonexpansiveness of the orthogonal projection operator. Simple algebraic rearrangement of the resulting inequality yields

$$\begin{aligned}2\langle H'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \le \frac{1}{t_k}(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2)\\ + t_k\|H'(\mathbf{x}_k)\|^2.\end{aligned} \tag{4.13}$$

On the other hand, by the strong convexity of the objective function H we have

$$H(\mathbf{x}^*) \ge H(\mathbf{x}_k) + \langle H'(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k\rangle + \frac{\sigma}{2}\|\mathbf{x}^* - \mathbf{x}_k\|^2,$$

and hence

$$\langle H'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \ge H(\mathbf{x}_k) - H(\mathbf{x}^*) + \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2,$$

which combined with (4.13) implies the inequality

$$\begin{aligned}H(\mathbf{x}_k) - H(\mathbf{x}^*) \le \frac{1}{2}\left(\frac{1}{t_k} - \sigma\right)\|\mathbf{x}_k - \mathbf{x}^*\|^2\\ - \frac{1}{2t_k}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \frac{t_k}{2}\|H'(\mathbf{x}_k)\|^2.\end{aligned}$$

Taking $t_k = \frac{1}{k\sigma}$ and summing over $k = 1, 2, \ldots, n$ we obtain

$$\begin{aligned}\sum_{k=1}^n (H(\mathbf{x}_k) - H^*) + \frac{n}{2}\|\mathbf{x}_{n+1} - \mathbf{x}^*\|^2 \le \frac{1}{2\sigma}\sum_{k=1}^n \frac{1}{k}\|H'(\mathbf{x}_k)\|^2\\ \le \frac{\gamma^2}{2\sigma}\sum_{k=1}^n \frac{1}{k} \le \frac{\gamma^2}{2\sigma}\ln(n+1).\end{aligned} \tag{4.14}$$

Thus,

$$\frac{\sum_{k=1}^n H(\mathbf{x}_k)}{n} - H(\mathbf{x}^*) \le \frac{\gamma^2}{2\sigma}\left(\frac{\ln(n+1)}{n}\right).$$

In addition, since $H(\mathbf{x}_k) \ge H^*$ for all k, it follows by (4.14) that for all $n \ge 1$

$$\frac{n}{2}\|\mathbf{x}_{n+1} - \mathbf{x}^*\|^2 \le \frac{\gamma^2}{2\sigma}\ln(n+1),$$

from which the inequality (4.12) follows a once. □

The above shows that under strong convexity of the objective, the rate of convergence of the subgradient projection can be improved from $O(1/\sqrt{k})$ to $O(\ln k/k)$. As was already mentioned, the subgradient projection method has two inherent disadvantages: it requires the computation of the orthogonal projection onto the feasible set X, and its convergence analysis assumes that the quantity γ must be finite. We will now show that it has a third disadvantage: its efficiency estimate is worse than the one of the dual proximal gradient method. We have seen that the rate of convergence of the sequence $\{\mathbf{x}_k\}$ generated by the fast dual proximal gradient method to $\mathbf{x}^*$ is $O(1/k)$. To compare the two methods, we need to consider the sequence of function values and its rate of convergence towards H^*. Since we want to look at a feasible point, we will consider the feasible sequence $\{P_X(\mathbf{x}_k)\}$ and establish the $O(1/k)$ rate of convergence of the fast dual proximal gradient method when applied on the direct primal formulation of the problem (P) as given in (4.10).

Theorem 4.3. *Let $\{\mathbf{y}_k\}$ be the sequence generated by the fast dual proximal gradient method with $L = \frac{\|\mathcal{A}\|^2}{\sigma}$ and let $\{\mathbf{x}_k\}$ be the corresponding primal sequence defined by (4.1). Then*

$$H(P_X(\mathbf{x}_k)) - H^* \le 2\gamma\frac{\|\mathcal{A}\|}{\sigma}\frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k}. \tag{4.15}$$

Proof. By the subgradient inequality, the Cauchy–Schwarz inequality and the nonexpansiveness of the projection operator, we have

$$\begin{aligned} H(P_X(\mathbf{x}_k)) - H(\mathbf{x}^*) &\le \langle H'(P_X(\mathbf{x}_k)), \mathbf{x}_k - \mathbf{x}^* \rangle \\ &\le \|H'(P_X(\mathbf{x}_k))\| \cdot \|P_X(\mathbf{x}_k) - \mathbf{x}^*\| \\ &\le \gamma \|\mathbf{x}_k - \mathbf{x}^*\| \\ &\le 2\gamma \frac{\|\mathcal{A}\|}{\sigma} \frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k} \quad (\text{by Theorem 4.1}). \quad \square \end{aligned}$$

To summarize, even when taking into account the strong convexity of the objective, the ergodic sequence of primal function values of the subgradient projection method converges at a rate of $O(\ln k/k)$ to the optimal value while the function values of the fast dual proximal gradient method converge with the superior rate of $O(1/k)$.

Acknowledgments

The work of Amir Beck partially supported by ISF grant #25312 and the work of Marc Teboulle was partially supported by ISF grant #99812.

References

[1] A. Auslender, M. Teboulle, Projected subgradient methods with non-Euclidean distances for non-differentiable convex minimization and variational inequalities, Math. Program. 120 (2009) 37–48.

[2] A. Beck, M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, Oper. Res. Lett. 31 (2003) 167–175.

[3] A. Beck, M. Teboulle, Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems, IEEE Trans. Image Process. 18 (2009) 2419–2434.

[4] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.

[5] D.P. Bertsekas, Constrained Optimization and Lagrangian Multipliers, Academic Press, New York, 1982.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.

[7] A. Chambolle, An algorithm for total variation minimization and applications, J. Math. Imaging Vision 20 (1–2) (2004) 89–97.

[8] P.L. Combettes, D. Dũng, B.C. Vũ, Dualization of signal recovery problems, Set-Valued Var. Anal. 18 (3–4) (2010) 373–404.

[9] P.L. Combettes, J.C. Presquet, Proximal splitting methods in signal processing, in: H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke, H. Wolkowicz (Eds.), Fixed-Point Algorithms for Inverse Problems in Science and Engineering, in: Springer Verlag series in Optimization and Its Applications, 2011, pp. 185–212.

[10] J.M. Fadili, G. Peyré, Total variation projection with first order schemes, IEEE Trans. Image Process. 20 (2011) 657–669.

[11] D. Gabay, Applications of the method of multipliers to variational inequalities, in: M. Fortin, R. Glowinski (Eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, Chapter IX, North-Holland, Amsterdam, 1983, pp. 299–340.

[12] R. Glowinski, P. Le Tallec, Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics, volume 9, in: Society for Industrial Mathematics, 1989.

[13] J.J. Moreau, Proximité et dualité dans un espace hilbertien, Bull. Soc. Math. France 93 (1965) 273–299.

[14] A.S. Nemirovsky, D.B. Yudin, Problem complexity and method efficiency in optimization, in: A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1983.

[15] Y. Nesterov, Gradient methods for minimizing composite objective function. 2007. CORE Report. Available at http://www.ecore.beDPs/dp1191313936.pdf.

[16] G.B. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert space, J. Math. Anal. Appl. 72 (2) (1979) 383–390.

[17] R.T. Rockafellar, Convex Analysis, Princeton Univ. Press, Princeton NJ, 1970.

[18] R.T. Rockafellar, Monotone operators and the proximal point algorithm, SIAM J. Control Optim. 14 (5) (1976) 877–898.

[19] R.T. Rockafellar, R.J.B Wets, Variational Analysis, in: Grundlehren der Mathematischen Wissenschaften, vol. 317, Springer-Verlag, Berlin, 1998.

[20] P. Tseng, Applications of a splitting algorithm to decomposition in convex programming and variational inequalities, SIAM J. Control Optim. 29 (1) (1991) 119–138.

[21] H. Uzawa, Iterative methods for concave programming, in: Studies in Linear and Nonlinear Programming, 1958, pp. 154–165.

[22] P. Weiss, L. Blanc-Féraud, G. Aubert, Efficient schemes for total variation minimization under constraints in image processing, SIAM J. Sci. Comput. 31 (2009) 2047–2080.