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

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

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 $C_1, C_2, \ldots, C_m \subseteq \mathbb{E}$ 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 $g(\mathbf{z}_1, \ldots, \mathbf{z}_m) = \sum_{i=1}^m \delta_{C_i}(\mathbf{z}_i)$ ($\delta_C(\cdot)$ being the indicator function of the set C). The linear operator $\mathcal{A} : \mathbb{E} \to \mathbb{E}^m$ is defined by $\mathcal{A}(\mathbf{x}) = \underbrace{(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})}_{m \text{ blocks}}$.

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

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

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 $u_j, j = 1, 2, \ldots, n$, 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 $x_j \in I_j, j = 1, 2, \ldots, n$, 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:

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

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

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

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