On the acceleration of the double smoothing technique for unconstrained convex optimization problems

Radu Ioan Boț* and Christopher Hendrich

Faculty of Mathematics, Chemnitz University of Technology,
D-09107 Chemnitz, Germany

(Received 3 May 2012; final version received 10 October 2012)

In this article, we investigate the possibilities of accelerating the double smoothing (DS) technique when solving unconstrained nondifferentiable convex optimization problems. This approach relies on the regularization in two steps of the Fenchel dual problem associated with the problem to be solved into an optimization problem having a differentiable strongly convex objective function with Lipschitz continuous gradient. The doubly regularized dual problem is then solved via a fast gradient method. The aim of this article is to show how the properties of the functions in the objective of the primal problem influence the implementation of the DS approach and its rate of convergence. The theoretical results are applied to linear inverse problems by making use of different regularization functionals.

Keywords: Fenchel duality; regularization; fast gradient method; image processing

AMS Subject Classifications: 90C25; 90C46; 47A52

1. Introduction

In this article, we develop an efficient algorithm based on the double smoothing (DS) approach for solving unconstrained nondifferentiable optimization problems of the type

\[
\begin{align*}
(P) \quad \inf_{x \in \mathcal{H}} \left\{ f(x) + g(Ax) \right\},
\end{align*}
\]

where $\mathcal{H}$ is a real Hilbert space, $f: \mathcal{H} \to \mathbb{R}$ and $g: \mathbb{R}^m \to \mathbb{R}$ are proper, convex and lower semicontinuous functions and $A: \mathcal{H} \to \mathbb{R}^m$ is a linear continuous operator fulfilling the feasibility condition $A(\text{dom} f) \cap \text{dom} g \neq \emptyset$. The DS technique for solving this class of optimization problems (see [8] for a fully finite-dimensional spaces version of it) assumes to efficiently solve the corresponding Fenchel dual problems and then to recover via an approximately optimal solution of the latter an approximately optimal solution of the primal. This technique, which represents a generalization of the approach developed in [10] for a special class of convex

*Corresponding author. Email: radu.bot@mathematik.tu-chemnitz.de

© 2012 Taylor & Francis
266  
R.I. Boţ and C. Hendrich

constrained optimization problems, makes use of the structure of the Fenchel dual and relies on the regularization of the latter in two steps into an optimization problem having a differentiable strongly convex objective function with Lipschitz continuous gradient. The regularized dual is then solved by a fast gradient method (FGM) which gives rise to a sequence of dual variables that solve the nonregularized dual problem after \( O\left(\frac{1}{\epsilon} \ln \left(\frac{1}{\epsilon}\right)\right) \) iterations, whenever \( f \) and \( g \) have bounded effective domains. In addition, the norm of the gradient of the regularized dual objective decreases by the same rate of convergence, a fact which is crucial in view of reconstructing an approximately optimal solution to \((P)\) after \( O\left(\frac{1}{\epsilon} \ln \left(\frac{1}{\epsilon}\right)\right) \) iterations [8].

The first aim of this article is to show that, whenever \( g \) is a strongly convex function, one can obtain the same convergence rate, even without imposing boundedness for its effective domain. Further we show that if, additionally, \( f \) is strongly convex or \( g \) is everywhere differentiable with a Lipschitz continuous gradient, then the convergence rate becomes \( O\left(\frac{1}{\epsilon} \ln \left(\frac{1}{\epsilon}\right)\right) \), while, if these supplementary assumptions are simultaneous fulfilled, then a convergence rate of \( O\left(\ln \left(\frac{1}{\epsilon}\right)\right) \) can be guaranteed.

The structure of this article is as follows. Section 2 is dedicated to some preliminaries on convex analysis and Fenchel duality. In Section 3, we employ the smoothing technique introduced in [12–14] in order to make the objective of the Fenchel dual problem of \((P)\) to be strongly convex and differentiable with Lipschitz continuous gradient. In Section 4, we solve the regularized dual problem via an efficient FGM, show how an approximately optimal primal solution can be recovered from a dual iterate and investigate the convergence properties of the sequence of primal optimal solutions. Section 5 addresses the question of how do the properties of the functions in the objective of \((P)\) influence the implementation of the DS approach and improve its rate of convergence. Finally, in Section 6, we consider an application of the presented approach in image deblurring and solve to this end by a linear inverse problem by using two different regularization functionals.

2. Preliminaries on convex analysis and Fenchel duality

Throughout this article \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle} \), respectively, denote the inner product and the norm of the real Hilbert space \( \mathcal{H} \), which is allowed to be infinite dimensional. The closure of a set \( C \subseteq \mathcal{H} \) is denoted by \( \text{cl}(C) \), while its indicator function is the function \( \delta_{C} : \mathcal{H} \to \mathbb{R} := \mathbb{R} \cup \{ \pm \infty \} \) defined by \( \delta_{C}(x) = 0 \) for \( x \in C \) and \( \delta_{C}(x) = +\infty \), otherwise. For a function \( f : \mathcal{H} \to \mathbb{R} \) we denote by \( \text{dom} f := \{ x \in \mathcal{H} : f(x) < +\infty \} \) its effective domain. We call \( f \) proper if \( \text{dom} f \neq \emptyset \) and \( f(x) > -\infty \) for all \( x \in \mathcal{H} \). The conjugate function of \( f \) is \( f^{\star} : \mathcal{H} \to \mathbb{R}^{+} \), \( f^{\star}(p) = \sup \{ \langle p, x \rangle - f(x) : x \in \mathcal{H} \} \) for all \( p \in \mathcal{H} \). The biconjugate function of \( f \) is \( f^{**} : \mathcal{H} \to \mathbb{R}^{+} \), \( f^{**}(x) = \sup \{ \langle x, p \rangle - f^{\star}(p) : p \in \mathcal{H} \} \) and, when \( f \) is proper, convex and lower semicontinuous, then, according to the Fenchel–Moreau theorem, one has \( f = f^{**} \). The (convex) subdifferential of the function \( f \) at \( x \in \mathcal{H} \) is the set \( \partial f(x) = \{ p \in \mathcal{H} : f(y) - f(x) \geq \langle p, y - x \rangle \text{ \forall } y \in \mathcal{H} \} \), if \( f(x) \in \mathbb{R} \), and is taken to be the empty set, otherwise.

Further, we consider the space \( \mathbb{R}^{m} \) endowed with the Euclidean inner product and norm, for which we use the same notations as for the real Hilbert space \( \mathcal{H} \), since no confusion can arise. By \( 1^{m} \) we denote the vector in \( \mathbb{R}^{m} \) with all entries equal to 1. For a subset \( C \) of \( \mathbb{R}^{m} \) we denote by \( \text{ri}(C) \) its relative interior, i.e. the interior of the set \( C \) relative to its affine hull. For a linear continuous operator \( A : \mathcal{H} \to \mathbb{R}^{m} \) the operator
$A^*: \mathbb{R}^m \rightarrow \mathcal{H}$, defined by $(A^* y, x) = (y, A x)$ for all $x \in \mathcal{H}$ and all $y \in \mathbb{R}^m$, is its so-called adjoint operator. By id: $\mathbb{R}^m \rightarrow \mathbb{R}^m$, id(x) = x, for all $x \in \mathbb{R}^m$ we denote the identity mapping on $\mathbb{R}^m$.

For a nonempty, convex and closed set $C \subseteq \mathcal{H}$ we consider the projection operator $\mathcal{P}_C: \mathcal{H} \rightarrow C$ defined as $x \mapsto \arg \min_{z \in C} \|x - z\|$. Having two functions $f, g: \mathcal{H} \rightarrow \mathbb{R}$, their infimal convolution is defined by $(f \square g)(x) = \inf_{y \in \mathcal{H}} (f(y) + g(x - y))$ for all $x \in \mathcal{H}$. The Moreau envelope of parameter $\gamma > 0$ of the function $f: \mathcal{H} \rightarrow \mathbb{R}$ is $\gamma f: \mathcal{H} \rightarrow \mathbb{R}$, defined as $\gamma f(x) := f \square \left(\frac{1}{2\gamma} \|\cdot\|^2\right)(x) = \inf_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\gamma} \|x - y\|^2 \right\}$ for all $x \in \mathcal{H}$.

The proximal point of $f$ at $x \in \mathbb{R}^n$ denotes the unique minimizer of the optimization problem

$$\inf_{y \in \mathbb{R}^m} \left\{ f(y) + \frac{1}{2} \|x - y\|^2 \right\}.$$ 

For $\rho > 0$ we say that the function $f: \mathcal{H} \rightarrow \mathbb{R}$ is $\rho$-strongly convex, if for all $x, y \in \mathcal{H}$ and all $\lambda \in (0, 1)$ it holds

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y) - \frac{\rho}{2} \lambda(1 - \lambda) \|x - y\|^2.$$ 

Notice that this is equivalent to saying that $x \mapsto f(x) - \frac{\rho}{2} \|x\|^2$ is convex.

For the optimization problem $(P)$ we consider the following standing assumptions: $f: \mathcal{H} \rightarrow \mathbb{R}$ is a proper, convex and lower semicontinuous function with a bounded effective domain, $g: \mathbb{R}^m \rightarrow \mathbb{R}$ is proper, $\mu$-strongly convex ($\mu > 0$) and lower semicontinuous function and $A: \mathcal{H} \rightarrow \mathbb{R}^m$ is a linear and continuous operator fulfilling $A(\text{dom } f) \cap \text{dom } g \neq \emptyset$.

**Remark 1** Different from the investigations made in [8] in a fully finite-dimensional setting, we strengthen here the convexity assumptions on $g$ (there $g$ was asked to be only proper, convex and lower semicontinuous), but allow in counterpart dom $g$ to be unbounded. The gain of weakening this assumption is emphasized by the applications considered in Section 6.

The Fenchel dual problem to $(P)$ (see, e.g. [5,6]) reads

$$(D) \quad \sup_{\rho \in \mathbb{R}^m} \left\{ -f^*(A^* \rho) - g^*(-\rho) \right\}. \quad (2)$$

We denote the optimal objective values of the optimization problems $(P)$ and $(D)$ by $\nu(P)$ and $\nu(D)$, respectively.

The conjugate functions of $f$ and $g$ can be written as

$$f^*(q) = \sup_{x \in \text{dom } f} \left\{ \langle q, x \rangle - f(x) \right\} = -\inf_{x \in \text{dom } f} \left\{ \langle -q, x \rangle + f(x) \right\} \quad \forall q \in \mathcal{H}$$

and

$$g^*(p) = \sup_{x \in \text{dom } g} \left\{ \langle p, x \rangle - g(x) \right\} = -\inf_{x \in \text{dom } g} \left\{ \langle -p, x \rangle + g(x) \right\} \quad \forall p \in \mathbb{R}^m,$$
respectively. According to [1, Theorem 11.9] and [4, Lemma 2.33], the optimization problems arising in the formulation of both $f^*(q)$ for all $q \in \mathcal{H}$ and $g^*(p)$ for all $p \in \mathbb{R}^m$ are solvable, fact which implies that $\text{dom } f^* = \mathcal{H}$ and $\text{dom } g^* = \mathbb{R}^m$, respectively.

By writing the dual problem $(D)$ equivalently as the infimum optimization problem

$$\inf_{p \in \mathbb{R}^m} \{ f^*(A^*p) + g^*(-p) \},$$

one can easily see that the Fenchel dual problem of the latter is

$$\sup_{x \in \mathcal{H}} \{-f^{**}(x) - g^{**}(Ax)\},$$

which, by the Fenchel–Moreau theorem, is nothing else than

$$\sup_{x \in \mathcal{H}} \{-f(x) - g(Ax)\}.$$

In order to guarantee strong duality for this primal-dual pair it is sufficient to ensure that (see, e.g. [5, Theorem 2.1]) $0 \in \text{ri}(A^*(\text{dom } g^*) + \text{dom } f^*)$. As $f^*$ has full domain, this regularity condition is automatically fulfilled, which means that $\nu(D) = \nu(P)$ and the primal optimization problem $(P)$ has an optimal solution. Due to the fact that $f$ and $g$ are proper and $A(\text{dom } f) \cap \text{dom } g \neq \emptyset$, this further implies $\nu(D) = \nu(P) \in \mathbb{R}$. Later we will assume that the dual problem $(D)$ has an optimal solution too, and that an upper bound of its norm is known.

Denote by $\theta: \mathbb{R}^m \to \mathbb{R}$, $\theta(p) = f^*(A^*p) + g^*(-p)$, the objective function of $(D)$. Hence, the dual can be equivalently written as

$$(D) \quad \Rightarrow \quad \inf_{p \in \mathbb{R}^m} \theta(p).$$

The assumptions made on $g$ yield that $p \mapsto g^*(-p)$ is differentiable and has a Lipschitz continuous gradient (see Section 3.1 for details). However, since in general one cannot guarantee the smoothness of $p \mapsto f^*(A^*p)$, the dual problem $(D)$ is a nondifferentiable convex optimization problem. Our goal is to solve this problem efficiently and to obtain from here an optimal solution to $(P)$. As in [8], we are overcoming the nonsatisfactory complexity of subgradient schemes, i.e. $O\left(\frac{1}{\epsilon}\right)$, by making use of smoothing techniques introduced in [12–14]. More precisely, we regularize first the objective function of $f^*(A^*p)$ by a quadratic term in order to obtain a smooth approximation of $p \mapsto f^*(A^*p)$. Then, we apply a second regularization to the new dual objective and minimize the regularized problem via an appropriate fast gradient scheme [8]. This will allow us to solve both optimization problems $(D)$ and $(P)$ approximately in $O\left(\frac{1}{\epsilon} \ln \left(\frac{1}{\epsilon}\right)\right)$ iterations. More than that, we will show that this rate of convergence can be improved when strengthening the assumptions imposed on $f$ and $g$.

### 3. The DS approach

#### 3.1. First smoothing

For a real number $\rho > 0$ the function $p \mapsto f^*(A^*p) = \sup_{x \in \mathcal{H}} \{ \langle A^*p, x \rangle - f(x) \}$ can be approximated by

$$f^*_\rho(A^*p) = \sup_{x \in \mathcal{H}} \left\{ \langle A^*p, x \rangle - f(x) - \frac{\rho}{2} \|x\|^2 \right\}.$$

(4)
For each \( p \in \mathbb{R}^m \) the maximization problem which occurs in the formulation of \( f^*_\rho(A^*p) \) has a unique solution (see, e.g. [4, Lemma 2.33]), fact which implies that \( f^*_\rho(A^*p) \in \mathbb{R} \).

For all \( p \in \mathbb{R}^m \) one can express the above regularization of the conjugate by means of the Moreau envelope of \( f \) as follows:

\[
-f^*_\rho(A^*p) = -\sup_{x \in \mathcal{H}} \left\{ \langle A^*p, x \rangle - f(x) - \frac{\rho}{2} \| x \|^2 \right\}
= \inf_{x \in \mathcal{H}} \left\{ f(x) + \frac{\rho}{2} \| A^*p - x \|^2 \right\} - \| A^*p \|^2 = \frac{i}{2} f \left( \frac{A^*p}{\rho} \right) - \frac{\| A^*p \|^2}{2\rho}.
\]

Consequently, one can transfer the differentiability properties of the Moreau envelope (see [1, Proposition 12.29]) to \( p \mapsto -(f^*_\rho \circ A^*)(p) \). For all \( p \in \mathbb{R}^m \) we have

\[
-\nabla(f^*_\rho \circ A^*)(p) = \frac{A}{\rho} \nabla \frac{i}{2} f \left( \frac{A^*p}{\rho} \right) - \frac{A}{\rho} A^*p = \frac{A}{\rho} \left( \frac{A^*p}{\rho} - x_{f,p} \right) = A x_{f,p},
\]

thus

\[
\nabla(f^*_\rho \circ A^*)(p) = A x_{f,p},
\]

where \( x_{f,p} \in \mathcal{H} \) is the proximal point of \( \frac{1}{\rho} f \) at \( \frac{A^*p}{\rho} \), namely the unique element in \( \mathcal{H} \) fulfilling [1, Proposition 12.29]

\[
\frac{i}{2} f \left( \frac{A^*p}{\rho} \right) = f(x_{f,p}) + \frac{\rho}{2} \| \frac{A^*p}{\rho} - x_{f,p} \|^2.
\]

By taking into account the nonexpansiveness of the proximal point mapping [1, Proposition 12.27], for \( p, q \in \mathbb{R}^m \) it holds

\[
\left\| \nabla(f^*_\rho \circ A^*)(p) - \nabla(f^*_\rho \circ A^*)(q) \right\| = \left\| A x_{f,p} - A x_{f,q} \right\| \leq \| A \| \| x_{f,p} - x_{f,q} \|
\leq \| A \| \left\| \frac{A^*p}{\rho} - \frac{A^*q}{\rho} \right\| \leq \frac{\| A \|^2}{\rho} \| p - q \|,
\]

Thus \( \frac{\| A \|^2}{\rho} \) is the Lipschitz constant of \( p \mapsto \nabla(f^*_\rho \circ A^*)(p) \).

Coming now to the function \( p \mapsto g^*(-p) = (g^* \circ -id)(p) \), let us notice first that, since \( g \) is proper, \( \mu \)-strongly convex and lower semicontinuous, \( g^* \) is differentiable and \( \nabla g^* \) is Lipschitz continuous with Lipschitz constant \( \frac{1}{\mu} \) (cf. [1, Theorem 18.15]). Thus \( (g^* \circ -id) \) is Fréchet differentiable too, and its gradient is Lipschitz continuous with Lipschitz constant \( \frac{1}{\mu} \). By denoting

\[
x_{g,p} := \nabla g^*(-p) = -\nabla(g^* \circ -id)(p),
\]

one has that \( -p \in \partial g(x_{g,p}) \) or, equivalently, \( 0 \in \partial \langle (p, \cdot) + g \rangle(x_{g,p}) \), which means that \( x_{g,p} \) is the unique optimal solution [4, Lemma 2.33] of the optimization problem

\[
\inf_{x \in \mathbb{R}^m} \{ \langle p, x \rangle + g(x) \}.
\]

Remark 2 If \( f \) is \( \rho \)-strongly convex \( (\rho > 0) \), then there is no need to apply the first regularization for \( p \mapsto f^*(A^*p) \), as this function is already Fréchet differentiable with a Lipschitz continuous gradient having a Lipschitz constant given by \( \frac{\| A \|^2}{\rho} \). Indeed, the
\( \rho \)-strong convexity of \( f \) implies that \( f^* \) is Fréchet differentiable with Lipschitz continuous gradient having a Lipschitz constant given by \( \frac{1}{\rho} \) \([1, \text{Theorem 18.15}]. \)

Hence, for all \( p, q \in \mathbb{R}^m \), we have

\[
\| \nabla (f^* \circ A^*)(p) - \nabla (f^* \circ A^*)(q) \| = \| A \nabla f^*(A^*p) - A \nabla f^*(A^*q) \| \\
\leq \frac{\| A \|}{\rho} \| A^*p - A^*q \| \leq \frac{\| A \|^2}{\rho} \| p - q \|.
\]

Taking

\[
x_{f,p} := \nabla f^*(A^*p),
\]

one has that \( 0 \in \partial (f - \langle A^*p, \cdot \rangle)(x_{f,p}) \), which means that \( x_{f,p} \) is the unique optimal solution \([4, \text{Lemma 2.33}]\) of the optimization problem

\[
\inf_{x \in H} \{ f(x) - \langle A^*p, x \rangle \}.
\]

By denoting \( D_f := \sup \left\{ \frac{\| x \|^2}{2} : x \in \text{dom} f \right\} \in \mathbb{R} \) we can relate \( f^* \circ A^* \) and its smooth approximation \( f^*_{\rho} \circ A^* \) as follows.

**Proposition 3** For all \( p \in \mathbb{R}^m \) it holds

\[
f^*_{\rho}(A^*p) \leq f^*(A^*p) \leq f^*_{\rho}(A^*p) + \rho D_f.
\]

Proof For \( p \in \mathbb{R}^m \) one has

\[
f^*_{\rho}(A^*p) = \langle A^*p, x_{f,p} \rangle - f(x_{f,p}) - \frac{\rho}{2} \| x_{f,p} \|^2 \leq \langle A^*p, x_{f,p} \rangle - f(x_{f,p}) \leq f^*(A^*p)
\]

\[
\leq \sup_{x \in \text{dom} f} \left\{ \frac{\| A^*p, x \|}{2} - f(x) - \frac{\rho}{2} \| x \|^2 \right\} + \sup_{x \in \text{dom} f} \left\{ \frac{\rho}{2} \| x \|^2 \right\}
\]

\[
= f^*_{\rho}(A^*p) + \rho D_f.
\]

For \( \rho > 0 \) let \( \theta_{\rho} : \mathbb{R}^m \rightarrow \mathbb{R} \) be defined by \( \theta_{\rho}(p) = f^*_{\rho}(A^*p) + g^*(-p) \). The function \( \theta_{\rho} \) is differentiable with a Lipschitz continuous gradient

\[
\nabla \theta_{\rho}(p) = \nabla (f^*_{\rho} \circ A^*)(p) + \nabla (g^* \circ -\text{id})(p) = A x_{f,p} - x_{g,p} \quad \forall p \in \mathbb{R}^m
\]

having Lipschitz constant \( L(\rho) := \frac{\| A \|^2}{\rho} + \frac{1}{\rho} \).

In consideration of Proposition 3 we get

\[
\theta_{\rho}(p) \leq \theta(p) \leq \theta_{\rho}(p) + \rho D_f \quad \forall p \in \mathbb{R}^m.
\]

In order to reconstruct an approximately optimal solution to the primal optimization problem \((P)\) it is not sufficient to ensure the convergence of \( \theta(\cdot) \) to \( -v(D) \), but we also need good convergence properties for the decrease of \( \| \nabla \theta_{\rho,\cdot}(\cdot) \| \) (cf. \([8,10]\)).

### 3.2. Second smoothing

In the following, a second regularization is applied to \( \theta_{\rho} \), as done in \([8,10]\), in order to make it strongly convex, fact which will allow us to use a fast gradient scheme with a
good convergence rate for the decrease of $\|\nabla\theta_\rho(\cdot)\|$. Therefore, adding the strongly convex function $\frac{\kappa}{2}\|\cdot\|^2$ to $\theta_\rho$, for some positive real number $\kappa$, gives rise to the following regularization of the objective function

$$\theta_{\rho,\kappa} : \mathbb{R}^m \rightarrow \mathbb{R}, \quad \theta_{\rho,\kappa}(p) := \theta_\rho(p) + \frac{\kappa}{2}\|p\|^2 = f^*_\rho(A^*p) + g^*(-p) + \frac{\kappa}{2}\|p\|^2,$$

which is obviously $\kappa$-strongly convex. We further deal with the optimization problem

$$\inf_{p \in \mathbb{R}^m} \theta_{\rho,\kappa}(p). \quad (6)$$

By taking into account [4, Lemma 2.33], the optimization problem (6) has a unique optimal solution, while the function $\theta_{\rho,\kappa}$ is differentiable and for all $p \in \mathbb{R}^m$ it holds

$$\nabla \theta_{\rho,\kappa}(p) = \nabla \left( \theta_\rho(\cdot) + \frac{\kappa}{2}\|\cdot\|^2 \right)(p) = Ax_\rho - x_{\rho,p} + \kappa p.$$

This gradient is Lipschitz continuous with constant $L(\rho, \kappa) := \frac{\|A\|^2}{\rho} + \frac{1}{\mu} + \kappa$.

**Remark 4** If $\theta_\rho$ is $\kappa$-strongly convex, then there is no need to apply the second regularization, as this function is already endowed with the properties of $\theta_{\rho,\kappa}$.

### 4. Solving the doubly regularized dual problem

#### 4.1. A fast gradient method

In the forthcoming sections, we denote by $p^*_\text{DS}$ the unique optimal solution of the optimization problem (6) and by $\theta^*_{\rho,\kappa} := \theta_{\rho,\kappa}(p^*_\text{DS})$ its optimal objective value. Further, we denote by $p^* \in \mathbb{R}^m$ an optimal solution to the dual optimization problem (D) and we assume that the upper bound

$$\|p^*\| \leq R \quad (7)$$

is available for some nonzero $R \in \mathbb{R}_+$. Furthermore, as in [8,10], we make use of the following FGM [11, Algorithm 2.2.11]

**Initialization:** Set $w_0 = p_0 := 0 \in \mathbb{R}^m$.

For $k \geq 0$: Set $p_{k+1} := w_k - \frac{1}{L(\rho, \kappa)}\nabla \theta_{\rho,\kappa}(w_k)$.

$$\text{(FGM)}$$

Set $w_{k+1} := p_{k+1} + \frac{\sqrt{L(\rho, \kappa) - \kappa}}{\sqrt{L(\rho, \kappa) + \kappa}}(p_{k+1} - p_k)$,

for minimizing the optimization problem (6), which has a strongly convex and differentiable optimization function with a Lipschitz continuous gradient. By taking into account [11, Theorem 2.2.3] we obtain a sequence $(p_k)_{k \geq 0} \subseteq \mathbb{R}^m$ satisfying

$$\theta_{\rho,\kappa}(p_k) - \theta_{\rho,\kappa}(p^*_\text{DS}) \leq \left( \theta_{\rho,\kappa}(0) - \theta_{\rho,\kappa}(p^*_\text{DS}) + \frac{\kappa}{2}\|p^*_\text{DS}\|^2 \right)e^{-k\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad (8)$$

$$= (\theta_\rho(0) - \theta_\rho(p^*_\text{DS}))e^{-k\sqrt{\frac{\kappa}{L(\rho, \kappa)}}} \quad \forall k \geq 0. \quad (9)$$
Since \( p^*_DS \) solves (6), we have \( \nabla \rho(p^*_DS) = 0 \) and, therefore [11, Theorem 2.1.5],
\[
\| \nabla \rho(p^*_DS) \|^2 \leq 2L(\rho, \kappa)(\theta(0) - \theta(p^*_DS)) e^{-k\sqrt{\frac{1}{\kappa \lambda}}}, \quad \forall k \geq 0. \tag{10}
\]
Due to the \( \kappa \)-strong convexity of \( \rho \), [11, Theorem 2.1.8] states
\[
\| p_k - p^*_DS \|^2 \leq \frac{2}{\kappa} \left( \theta(p_k) - \theta(p^*_DS) \right)
\leq \frac{2}{\kappa} (\theta(0) - \theta(p^*_DS)) e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}}, \quad \forall k \geq 0. \tag{11}
\]
We first prove that the rates of convergence for the decrease of \( \theta(p_k) - \theta(p^*) \) and \( \| \nabla \rho(p_k) \| \) coincide, being equal to \( O\left( e^{\frac{1}{2} \ln \left( \frac{1}{\lambda} \right)} \right) \), and that they can be improved when \( f \) and/or \( g \) fulfill additional assumptions. We also show how \( \epsilon \)-optimal solutions to the primal problem \( (P) \) can be recovered from the sequence of dual variables \( (p_k)_{k \geq 0} \). To this aim we will act in the lines of the considerations from [8,10] and this is why we refer the reader to these papers for detailed argumentations in this sense.

4.2. Convergence of \( \theta(p_k) \) to \( \theta(p^*) \)

Using again [11, Theorem 2.1.8] we obtain
\[
\| p^*_DS \|^2 \leq \frac{2}{\kappa} \left( \theta(0) - \theta(p^*_DS) \right)
= \frac{2}{\kappa} \left( \theta(0) - \theta(p^*_DS) - \frac{\kappa}{2} \| p^*_DS \|^2 \right),
\]
which implies that
\[
\| p^*_DS \|^2 \leq \frac{1}{\kappa} (\theta(0) - \theta(p^*_DS)). \tag{12}
\]
In order to estimate the function values, we notice that formula (9) states
\[
\theta(p_k) - \theta(p^*_DS) \leq (\theta(0) - \theta(p^*_DS)) e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}} + \frac{\kappa}{2} (\| p^*_DS \|^2 - \| p_k \|^2), \quad \forall k \geq 0.
\]
The last term in the inequality above can be estimated via
\[
\| p^*_DS \|^2 - \| p_k \|^2 \leq \| p^*_DS - p_k \| (2 \| p^*_DS \| + \| p_k - p^*_DS \|)
\leq \frac{2}{\kappa} (\theta(0) - \theta(p^*_DS)) e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}} + 2 \sqrt{\frac{\kappa}{2}} (\theta(0) - \theta(p^*_DS)) e^{\frac{-\kappa}{\sqrt{\frac{1}{\kappa \lambda}}}}
\leq \frac{2 + \sqrt{\frac{\kappa}{2}}}{\kappa} (\theta(0) - \theta(p^*_DS)) e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}}.
\]
Thus we obtain for all \( k \geq 0 \)
\[
\theta(p_k) - \theta(p^*_DS) \leq (\theta(0) - \theta(p^*_DS)) \left( e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}} + (1 + \sqrt{\frac{\kappa}{2}}) e^{\frac{-\kappa}{\sqrt{\frac{1}{\kappa \lambda}}}} \right)
\leq (2 + \sqrt{2}) (\theta(0) - \theta(p^*_DS)) e^{\frac{-k}{\sqrt{\frac{1}{\kappa \lambda}}}}. \tag{13}
\]
Further, we have \( \theta(0) \leq \theta(0), \theta(p^*_DS) \geq \theta(p^*_DS) - \rho D_f \geq \theta(p^*) - \rho D_f \) and, from here,
\[
\theta(p_k) - \theta(p^*_DS) \leq \theta(0) - \theta(p^*) + \rho D_f. \tag{14}
\]
Hence, using (5),
\[ \theta_\rho(p^*_{DS}) \leq \theta_\rho(p^*_{DS}) + \frac{\kappa}{2} \| p^*_{DS} \|^2 \leq \theta_\rho(p^*) + \frac{\kappa}{2} \| p^* \|^2 \]
and from here it follows for all \( k \geq 0 \)
\[ \theta(p_k) - \theta(p^*) \leq \rho D_f + \frac{\kappa}{2} \| p^* \|^2 + \theta_\rho(p_k) - \theta_\rho(p^*) \leq \rho D_f + \frac{\kappa}{2} \| p^* \|^2 + \theta_\rho(p_k) - \theta_\rho(p^*) \]
\[ \leq (13) \rho D_f + \frac{\kappa}{2} R^2 + (2 + \sqrt{2})(\theta(0) - \theta(p^*) + \rho D_f) e^{-\frac{\kappa}{2^2 \sqrt{\frac{R^2}{3}}}}. \] (15)

Next we fix \( \varepsilon > 0 \). In order to get \( \theta(p_k) - \theta(p^*) \leq \varepsilon \) after a certain amount of iterations \( k \), we force all three terms in (15) to be less than or equal to \( \frac{\varepsilon}{2} \). To this end we choose first
\[ \rho := \rho(\varepsilon) = \frac{\varepsilon}{3D_f} \quad \text{and} \quad \kappa := \kappa(\varepsilon) = \frac{2\varepsilon}{3R^2}. \] (16)

With these new parameters we can simplify (15) to
\[ \theta(p_k) - \theta(p^*) \leq \frac{2\varepsilon}{3} + (2 + \sqrt{2})(\theta(0) - \theta(p^*) + \frac{\varepsilon}{3}) e^{-\frac{\kappa}{2^2 \sqrt{\frac{R^2}{3}}}} \forall k \geq 0, \]
thus, the second term in the expression on the right-hand side of the above estimate determines the number of iterations needed to obtain \( \varepsilon \)-accuracy for the dual objective function \( \theta \). Indeed, we have
\[ \frac{\varepsilon}{3} \geq (2 + \sqrt{2})(\theta(0) - \theta(p^*) + \frac{\varepsilon}{3}) e^{-\frac{\kappa}{2^2 \sqrt{\frac{R^2}{3}}}} \]

\[ \Leftrightarrow k \geq 2 \sqrt{\frac{3L(\rho, \kappa)}{\kappa}} \ln \left( \frac{3(2 + \sqrt{2})(\theta(0) - \theta(p^*) + \frac{\varepsilon}{3})}{\epsilon} \right). \] (17)

Noticing that
\[ \frac{L(\rho, \kappa)}{\kappa} = \frac{\| A \|^2}{\rho \kappa} + \frac{1}{\mu k} + 1 = \frac{1}{\epsilon^2} \left( \frac{9\| A \|^2 D_f R^2}{2} + \frac{3R^2 \epsilon}{2\mu} + \epsilon^2 \right), \]
in order to obtain an approximately optimal solution to (D), we need \( k = O(\frac{1}{\epsilon} \ln(\frac{1}{\epsilon})) \) iterations.

### 4.3. Convergence of \( \| \nabla \theta_\rho(p_k) \| \) to 0

Guaranteeing \( \epsilon \)-optimality for the objective value of the dual is not sufficient for solving the primal optimization problem with a good convergence rate, as we need at least the same convergence rate for the decrease of \( \| \nabla \theta_\rho(p_k) \| = \| A x_{f,p_k} - x_{k,p_k} \| \) to 0 in order to ensure primal feasibility. Within this section we show that this is actually the case (see also [10]). It holds
\[ \| \nabla \theta_\rho(p_k) \| \leq \| \nabla \theta_\rho(p_k) - \kappa p_k \| \leq \| \nabla \theta_\rho(p_k) \| + \kappa \| p_k \| \forall k \geq 0. \]
The first term on the right-hand side of the above equation can be estimated using (10), namely
\[ \| \nabla \theta_{\rho,k}(p_k) \| \leq \sqrt{2L(\rho, \kappa)(\theta(0) - \theta(p_{DS}^*))} e^{-\frac{\epsilon}{\sqrt{2L(\rho, \kappa)}}} \quad \forall k \geq 0, \]
while for the second term, we use
\[ \| p_k \| = \| p_k - p_{DS}^* + p_{DS}^* \| \leq \| p_k - p_{DS}^* \| + \| p_{DS}^* \| \]
\[ \leq \frac{1}{\kappa} \left( \theta(0) - \theta(p_{DS}^*) \right) e^{-\frac{\epsilon}{\sqrt{2L(\rho, \kappa)}}} + \| p_{DS}^* \| \quad \forall k \geq 0. \quad (18) \]

Moreover, we notice that
\[ \theta(p^*) + \kappa \frac{D_f}{\kappa} p^* \leq \theta(p_{DS}^*) - \rho D_f + \frac{\kappa}{2} \| p_{DS}^* \|^2 \geq \theta(p^*) - \rho D_f + \frac{\kappa}{2} \| p_{DS}^* \|^2, \]
which implies that \( \| p_{DS}^* \|^2 \leq \| p^* \|^2 + \frac{2\rho D_f}{\kappa}. \) Hence,
\[ \| p_{DS}^* \| \leq \sqrt{\| p^* \|^2 + \frac{2\rho D_f}{\kappa}} = \sqrt{\| p^* \|^2 + \frac{2\rho D_f}{3\kappa}} = \sqrt{\| p^* \|^2 + R^2} \leq \sqrt{2R}, \quad (19) \]
which, combined with the previous estimates, (14) and (16), provides for all \( k \geq 0 \)
\[ \| \nabla \theta_{\rho}(p_k) \| \leq \left( \sqrt{L(\rho, \kappa)} + \sqrt{\kappa} \right) \sqrt{2(\theta(0) - \theta(p_{DS}^*))} e^{-\frac{\epsilon}{\sqrt{2L(\rho, \kappa)}}} + \sqrt{2\kappa R} \]
\[ \leq \left( \sqrt{L(\rho, \kappa)} + \sqrt{\kappa} \right) \sqrt{2(\theta(0) - \theta(p^*) + \frac{\epsilon}{3})} e^{-\frac{\epsilon}{\sqrt{2L(\rho, \kappa)}}} + \frac{2\sqrt{2\epsilon}}{3R}. \quad (20) \]

For \( \epsilon > 0 \) fixed, the first term in (20) decreases by the iteration counter \( k \), and, in order to ensure \( \| \nabla \theta_{\rho}(p_k) \| \leq \frac{\epsilon}{R} \), we need
\[ k \geq 2 \sqrt{\frac{L(\rho, \kappa)}{\kappa}} \ln \left( \frac{3R \left( \sqrt{L(\rho, \kappa)} + \sqrt{\kappa} \right) \sqrt{2(\theta(0) - \theta(p^*) + \frac{\epsilon}{3})}}{(3 - 2\sqrt{2})\epsilon} \right) \quad (21) \]
iteration steps. Summarizing, by taking into account (16), we can ensure
\[ \theta(p_k) - \theta(p^*) \leq \epsilon \quad \text{and} \quad \| \nabla \theta_{\rho}(p_k) \| \leq \frac{\epsilon}{R} \quad (22) \]
in \( k = O(\sqrt{\ln \frac{1}{\epsilon}}) \) iterations.

4.4. Constructing an approximate primal solution

Since our main focus is to solve the primal optimization problem \( (P) \), we prove as follows that the sequences \( (x_{f,p})_{k \geq 0} \subseteq \text{dom} f \) and \( (x_{g,p})_{k \geq 0} \subseteq \text{dom} g \) constructed in Section 3.1 contain all the information one needs to recover approximately optimal solutions to \( (P) \) (see [8,10] for a similar approach). Let \( k := k(\epsilon) \) be the smallest index satisfying (17) and (21), thus guaranteeing (22).

Since \( \theta_{\rho}(p_k) - \theta(p^*) \leq \theta(p_k) - \theta(p^*) \leq \epsilon \) and
\[ \theta_{\rho}(p_k) - \theta(p^*) \leq \theta(p_k) - \rho D_f - \theta(p^*) \leq \theta(p_k) - \theta(p^*) - \frac{\epsilon}{3} \leq -\frac{\epsilon}{3}, \]

respectively.
it holds \(|\theta(p_k) - \theta(p^*)| \leq \epsilon\) for all \(k \geq 0\). Further, we have
\[
\theta(p_k) = f^*(A^*p_k) + g^*(-p_k) = \langle p_k, Ax_{f,p_k} \rangle - f(x_{f,p_k}) - \frac{\rho}{2} \|x_{f,p_k}\|^2 - \langle p_k, x_{g,p_k} \rangle - g(x_{g,p_k})
\]
and from here (notice that \(-v(D) = \theta(p^*)\))
\[
f(x_{f,p_k}) + g(x_{g,p_k}) - v(D) = \langle p_k, \nabla \theta_{\rho}(p_k) \rangle + (\theta(p^*) - \theta(p_k)) - \frac{\rho}{2} \|x_{f,p_k}\|^2 \quad \forall k \geq 0.
\]

It follows
\[
|f(x_{f,p_k}) + g(x_{g,p_k}) - v(D)| \leq \|p_k\| \|\nabla \theta_{\rho}(p_k)\| + |\theta(p^*) - \theta(p_k)| + \frac{\rho}{2} \|x_{f,p_k}\|^2
\leq \|p_k\| \|\nabla \theta_{\rho}(p_k)\| + \epsilon + \rho D_f
\leq \frac{\rho}{2} \|x_{f,p_k}\|^2 + 2\epsilon \leq \frac{\epsilon}{R} \|p_k\| + 2\epsilon \quad \forall k \geq 0.
\]

In the light of (18) and (19), it holds
\[
\|p_k\| \leq \sqrt{\frac{2}{\kappa}} \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3}\right) e^{\frac{\epsilon}{2} \sqrt{\frac{1}{\kappa R} \epsilon}} + \sqrt{2R}
\leq R \sqrt{\frac{3}{\epsilon}} \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3}\right) e^{\frac{\epsilon}{2} \sqrt{\frac{1}{\kappa R} \epsilon}} + \sqrt{2R}.
\]
Finally, we obtain
\[
|f(x_{f,p_k}) + g(x_{g,p_k}) - v(D)| \leq \sqrt{3\epsilon \left(\theta(0) - \theta(p^*) + \frac{\epsilon}{3}\right) e^{\frac{\epsilon}{2} \sqrt{\frac{1}{\kappa R} \epsilon}}} + (2 + \sqrt{2})\epsilon,
\]
which, due to the choice of \(k = k(\epsilon)\), fulfils
\[
|f(x_{f,p_k}) + g(x_{g,p_k}) - v(D)| \leq 5\epsilon.
\]

By taking into account weak duality, i.e. \(v(D) \leq v(P)\), we conclude that \(x_{f,p_k} \in \text{dom } f\) and \(x_{g,p_k} \in \text{dom } g\) can be seen as approximately optimal solutions to \((P)\).

\subsection*{4.5. Existence of an optimal solution}

We close this section by a convergence analysis on the two sequences of primal approximate optimal solutions when \(\epsilon\) converges to zero. To this end let \((\epsilon_n)_{n \geq 0} \subseteq \mathbb{R}_+\) be a decreasing sequence of positive scalars with \(\lim_{n \to \infty} \epsilon_n = 0\). For each \(n \geq 0\), the DS algorithm (FGM) with smoothing parameters \(\rho_n\) and \(k\epsilon_n\) given by (16) requires at least \(k = k(\epsilon_n)\) iterations to fulfill (17) and (21). For \(n \geq 0\) we denote
\[
\overline{x}_n := x_{f,p_k(\epsilon_n)} \in \text{dom } f \quad \text{and} \quad \overline{y}_n := x_{g,p_k(\epsilon_n)} \in \text{dom } g.
\]

Due to the boundedness of \(\text{dom } f\), its closure \(\text{cl}(\text{dom } f)\) is weakly compact [1, Theorem 3.3] and there exists a subsequence \((\overline{x}_n)_{n \geq 0}\) and \(\overline{x} \in \mathcal{H}\) such that \(\overline{x}_n\) weakly converges to \(\overline{x} \in \text{cl}(\text{dom } f)\) when \(l \to +\infty\). Since \(A : \mathcal{H} \to \mathbb{R}^m\) is linear and
continuous, the sequence \( A\overline{n} \) will converge to \( A\overline{x} \) when \( l \to +\infty \). In view of relation (22) we get
\[
0 \leq \| A\overline{n} - \overline{y}_n \| \leq \frac{\epsilon_n}{R} \quad \forall l \geq 0. \tag{24}
\]
This means that the sequence \( (\overline{y}_n)_{l \geq 0} \subseteq \text{dom} \, g \) is obviously bounded, hence there exists a subsequence of it (still denoted by \( (\overline{y}_n)_{l \geq 0} \)) and an element \( \overline{y} \in \text{cl} \,(\text{dom} \, g) \) such that \( \overline{y}_n \to \overline{y} \) when \( l \to +\infty \). Taking \( l \to +\infty \) in (24) it follows \( A\overline{x} = \overline{y} \). Furthermore, due to (23), we have
\[
f(\overline{x}_n) + g(\overline{y}_n) \leq v(D) + 5\epsilon_n \quad \forall l \geq 0
\]
and, by using the lower semicontinuity of \( f \) and \( g \) and [1, Theorem 9.1], we obtain
\[
f(\overline{x}) + g(A\overline{x}) \leq \liminf_{l \to \infty} \left\{ f(\overline{x}_n) + g(\overline{y}_n) \right\}
\leq \lim_{l \to \infty} \left\{ v(D) + 5\epsilon_n \right\} = v(D) \leq v(P).
\]
Since \( v(P) \in \mathbb{R} \), we have \( \overline{x} \in \text{dom} \, f \) and \( A\overline{x} \in \text{dom} \, g \), which yields that \( \overline{x} \) is an optimal solution to \( (P) \).

5. Improving the convergence rates

In this section, we investigate how additional assumptions on the functions \( f \) and/or \( g \) influence the implementation of the DS approach, its rate of convergence and eventually allow a weakening of the standing assumptions made in this article. In all the three situations addressed here the construction of the approximate primal solutions and the proof of the existence of an optimal solution to the primal problem can be made in analogy to the Sections 4.4 and 4.5, respectively. It is worth noting that the additional assumptions furnish an improvement of the complexity, which is motivated by the fact that constants of strong convexity and/or Lipschitz constants of the gradient are already available, thus they do not need to be in the smoothing process constructed as functions of the level of accuracy \( \epsilon \).

5.1. The case \( f \) is strongly convex

In addition to the standing assumptions, we assume first that the function \( f : \mathcal{H} \to \mathbb{R} \) is \( \rho \)-strongly convex (\( \rho > 0 \)), but remove the boundedness assumption on its domain. In this situation the first smoothing, as done in Section 3.1, can be omitted and the FGM can be applied to the minimization problem
\[
\inf_{p \in \mathbb{R}^m} \theta_\kappa(p), \tag{25}
\]
where \( \theta_\kappa : \mathbb{R}^m \to \mathbb{R} \), \( \theta_\kappa := f^*(A^*p) + g^*(-p) + \frac{k}{2} \| p \|^2 \), with \( \kappa > 0 \), is a \( \kappa \)-strongly convex and differentiable function with Lipschitz continuous gradient. The Lipschitz constant of \( \nabla \theta_\kappa \) is \( L(\kappa) := \frac{\| A \|^2}{\rho} + \frac{1}{\mu} + \kappa \).
This gives rise to a sequence \((p_k)_{k \geq 0}\) satisfying
\[
\theta_k(p_k) - \theta_k(p_{DS}^*) \leq \left(\theta_k(0) - \theta_k(p_{DS}^*) + \frac{\kappa}{2} \| p_{DS}^* \|^2\right)e^{-k\sqrt{\frac{\kappa}{2}}} \tag{26}
\]
\[
= (\theta(0) - \theta(p_{DS}^*))e^{-k\sqrt{\frac{\kappa}{2}}} \quad \forall k \geq 0,
\tag{27}
\]
where \(p_{DS}^*\) denotes the unique optimal solution of the problem (25). Thus, from (27) it follows
\[
\| \nabla \theta_k(p_k) \|^2 \leq 2L(\kappa)(\theta(0) - \theta(p_{DS}^*))e^{-k\sqrt{\frac{\kappa}{2}}} \tag{28}
\]
and
\[
\| p_k - p_{DS}^* \|^2 \leq \frac{2}{\kappa}(\theta_k(p_k) - \theta_k(p_{DS}^*)) \leq \frac{2}{\kappa}(\theta(0) - \theta(p_{DS}^*))e^{-k\sqrt{\frac{\kappa}{2}}} \quad \forall k \geq 0. \tag{29}
\]
Additionally, in all iterations \(k \geq 0\) we have
\[
\| p_{DS}^* \|^2 \leq \frac{1}{\kappa}(\theta(0) - \theta(p_{DS}^*)) \tag{30}
\]
and
\[
\| p_{DS}^* \|^2 - \| p_k \|^2 \leq \| p_k - p_{DS}^* \| (2\| p_{DS}^* \| + \| p_k - p_{DS}^* \|) \leq \frac{2 + 2\sqrt{2}}{\kappa} \theta(0) - \theta(p_{DS}^*)e^{-\frac{\kappa}{4}\sqrt{\frac{\kappa}{2}}},
\]
thus
\[
\theta(p_k) - \theta(p_{DS}^*) \leq \left(\theta(0) - \theta(p_{DS}^*)\right)e^{-k\sqrt{\frac{\kappa}{2}}} + \frac{\kappa}{2}\left(\| p_{DS}^* \|^2 - \| p_k \|^2\right) \leq \left(\theta(0) - \theta(p_{DS}^*)\right)e^{-k\sqrt{\frac{\kappa}{2}} + (1 + \sqrt{2})e^{-\frac{\kappa}{4}\sqrt{\frac{\kappa}{2}}}} \leq (2 + \sqrt{2})(\theta(0) - \theta(p_{DS}^*))e^{-\frac{\kappa}{8}\sqrt{\frac{\kappa}{2}}} \quad \forall k \geq 0.
\]
We denote by \(p^* \in \mathbb{R}^n\) an optimal solution to the dual optimization problem (D) and assume that the upper bound \(\| p^* \| \leq R\) is available for some nonzero \(R \in \mathbb{R}_+\). Thus, since \(\theta(p_{DS}^*) \leq \theta_k(p_{DS}^*) \leq \theta_k(p^*) = \theta(p^*) + \frac{\kappa}{2} \| p^* \|^2\), we obtain for all \(k \geq 0\)
\[
\theta(p_k) - \theta(p^*) \leq \frac{\kappa}{2} \| p^* \|^2 + \theta(p_k) - \theta(p_{DS}^*) \leq \frac{\kappa}{2} R^2 + (2 + \sqrt{2})(\theta(0) - \theta(p^*))e^{-\frac{\kappa}{8}\sqrt{\frac{\kappa}{2}}}.
\]
Hence, when \(\epsilon > 0\), in order to guarantee \(\epsilon\)-accuracy for the dual objective function we can force both terms in the above estimate to be less than or equal to \(\frac{\epsilon}{2}\). Thus, by taking
\[
\kappa := \kappa(\epsilon) = \frac{\epsilon}{R^2},
\]
this time we will need to this end, in contrast to (17),
\[ k \geq 2\sqrt{\frac{L(\kappa)}{\kappa}} \ln \left( \frac{2(2 + \sqrt{2})(\theta(0) - \theta(p^*))}{\epsilon} \right), \]
i.e. \( k = O\left( \frac{1}{\sqrt{\epsilon}} \ln \left( \frac{1}{\epsilon} \right) \right) \) iterations. Further, using (28) we have
\[ \| \nabla \theta_k(p_k) \| \leq \sqrt{2L(\kappa)(\theta(0) - \theta(p^*))} e^{-\frac{k}{\sqrt{\kappa}}} + \kappa R, \]
on the other hand, using
\[ \| p_k \| \leq \| p_k - p^*_D \| + \| p^*_D \| \leq \frac{2}{\kappa} (\theta(0) - \theta(p^*)) e^{-\frac{k}{\sqrt{\kappa}}} + \kappa R, \]
and the relation \( \theta(p^*) + \frac{\kappa}{\epsilon} \| p^*_D \|^2 \leq \theta_k(p^*_D) \leq \theta_k(p^*) = \theta(p^*) + \frac{\kappa}{\epsilon} \| p^* \|^2 \), which yields \( \| p^*_D \| \leq \| p^* \| \leq R \), we obtain
\[ \| \nabla \theta(p_k) \| \leq \| \nabla \theta_k(p_k) \| + \kappa \| p_k \| \leq \left( \sqrt{L(\kappa) + \kappa} \right) \sqrt{2(\theta(0) - \theta(p^*))} e^{-\frac{k}{\sqrt{\kappa}}} + \kappa R \]
\[ = \left( \sqrt{L(\kappa) + \kappa} \right) \sqrt{2(\theta(0) - \theta(p^*))} e^{-\frac{k}{\sqrt{\kappa}}} + e R, \quad \forall k \geq 0. \]
Therefore, in order to guarantee \( \| Ax_{f,p_k} - x_{g,p_k} \| = \| \nabla \theta(p_k) \| \leq \frac{\epsilon}{R} \), we need \( k = O\left( \frac{1}{\sqrt{\epsilon}} \ln \left( \frac{1}{\epsilon} \right) \right) \) iterations, which coincides with the convergence rate for the dual objective values.

5.2. The case \( g \) is everywhere differentiable with Lipschitz continuous gradient
Throughout this section, additionally to the standing assumptions, we assume that \( g: \mathbb{R}^m \to \mathbb{R} \) has full domain and is differentiable with \( \frac{1}{\kappa} \)-Lipschitz continuous gradient, for \( \kappa > 0 \). In this situation the second smoothing, as done in Section 3.2, can be omitted and the FGM can be applied to the minimization problem
\[ \inf_{p \in \mathbb{R}^m} \theta_\rho(p), \quad (31) \]
where \( \theta_\rho: \mathbb{R}^m \to \mathbb{R}, \theta_\rho := f^*_\rho(A^*p) + g^*(-p) \), is \( \kappa \)-strongly convex due to [1, Theorem 18.15] and differentiable with Lipschitz continuous gradient. The Lipschitz constant of \( \nabla \theta_\rho \) is \( L(\rho) := \| A \|^2 + \frac{1}{\rho} \).
This gives rise to a sequence \( (p_k)_{k \geq 0} \) satisfying
\[ \theta_\rho(p_k) - \theta_\rho(p^*_D) \leq \left( \theta_\rho(0) - \theta_\rho(p^*_D) + \frac{\kappa}{2} \| p^*_D \|^2 \right) e^{-k \sqrt{\pi n}}, \quad (32) \]
\[ \leq 2(\theta_\rho(0) - \theta_\rho(p^*_D)) e^{-k \sqrt{\pi n}}, \quad (33) \]
and
\[ \| \nabla \theta_\rho(p_k) \|^2 \leq 4L(\rho)(\theta_\rho(0) - \theta_\rho(p^*_D)) e^{-k \sqrt{\pi n}} \quad \forall k \geq 0, \quad (34) \]
where \( p^*_D \) denotes the unique optimal solution of the problem (31). We denote by \( p^* \in \mathbb{R}^m \) the unique optimal solution of the dual optimization problem \( (D) \) and would
like to notice that in this context it is not necessary to know an upper bound of the norm of the dual optimal solution.

Since \( \theta_{\rho}(0) \leq \theta(0) \) and \( \theta_{\rho}(p_{DS}^*) \geq \theta(p_{DS}^*) - \rho D_f \geq \theta(p^*) - \rho D_f \), we obtain

\[
\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*) \leq \theta(0) - \theta(p^*) +\rho D_f.
\]

(35)

On the other hand, since \( \theta_{\rho}(p_k) - \theta_{\rho}(p_{DS}^*) \geq \theta(p_k) - \rho D_f - \theta(p^*) \), it follows

\[
\theta(p_k) - \theta(p^*) \leq \rho D_f + \theta(p_k) - \theta(p_{DS}^*) \leq \rho D_f + \rho D_f e^{-k \sqrt{\frac{\rho}{D_f}}} \forall k \geq 0.
\]

Hence, when \( \epsilon > 0 \), in order to guarantee \( \epsilon \)-optimality for the dual objective, we force both terms in the above estimate less than or equal to \( \frac{\epsilon}{2} \). By taking

\[
\rho := \rho(\epsilon) = \frac{\epsilon}{2D_f},
\]

in contrast to (17), we need

\[
k \geq \sqrt{\frac{L(\rho)}{\kappa}} \ln \left( \frac{4(\theta(0) - \theta(p^*) + \frac{\epsilon}{2})}{\epsilon} \right),
\]

i.e. \( k = O\left(\frac{1}{\epsilon} \ln(\frac{1}{2})\right) \) iterations to obtain \( \epsilon \)-accuracy for the dual objective values.

From (34) we obtain as well

\[
\| \nabla \theta_{\rho}(p_k) \| \leq 2\sqrt{L(\rho)(\theta_{\rho}(0) - \theta_{\rho}(p_{DS}^*))} e^{-\frac{k}{2}\sqrt{\frac{\rho}{D_f}}}
\leq 2\sqrt{L(\rho)(\theta(0) - \theta(p^*) + \rho D_f)} e^{-\frac{k}{2}\sqrt{\frac{\rho}{D_f}}}
\leq 2\sqrt{L(\rho)(\theta(0) - \theta(p^*) + \frac{\epsilon}{2})} e^{-\frac{k}{2}\sqrt{\frac{\rho}{D_f}}} \forall k \geq 0.
\]

Therefore, in order to guarantee \( \| A x_{f,p_k} - x_{g,p_k} \| = \| \nabla \theta_{\rho}(p_k) \| \leq \epsilon \), we need \( k = O\left(\frac{1}{\epsilon^2} \ln(\frac{1}{2})\right) \) iterations, which is the same convergence rate as for the dual objective values.

5.3. The case \( f \) is strongly convex and \( g \) is everywhere differentiable with Lipschitz continuous gradient

The third favourable situation which we address is when, additionally to the standing assumptions, the function \( f: \mathcal{H} \to \mathbb{R} \) is \( \rho \)-strongly convex \( (\rho > 0) \), however without assuming anymore that \( \text{dom} f \) is bounded, and the function \( g: \mathbb{R}^m \to \mathbb{R} \) has full domain and is differentiable with \( \frac{1}{\kappa} \)-Lipschitz continuous gradient \( (\kappa > 0) \). In this case both the first and second smoothing can be omitted and the FGM can be applied to the minimization problem

\[
\inf_{\rho \in \mathbb{R}^m} \theta(p),
\]

where \( \theta: \mathbb{R}^m \to \mathbb{R}, \theta := f^*(A^*p) + g^*(-p), \) is a \( \kappa \)-strongly convex and differentiable function with Lipschitz continuous gradient. The Lipschitz constant of \( \nabla \theta \) is
L := \frac{\|A\|^2}{\rho} + \frac{1}{\mu}. We denote by \( p^* \in \mathbb{R}^n \) the unique optimal solution of (D), for which it is not necessary to know an upper bound of its norm.

This gives rise to a sequence \((p_k)_{k \geq 0}\) satisfying

\[
\theta(p_k) - \theta(p^*) \leq (\theta(0) - \theta(p^*) + \frac{\kappa}{2} \|p^*\|^2) e^{-k\sqrt{\kappa}} \leq 2(\theta(0) - \theta(p^*)) e^{-k\sqrt{\kappa}}
\]

and

\[
\|\nabla \theta(p_k)\|^2 \leq 4L(\theta(0) - \theta(p^*)) e^{-k\sqrt{\kappa}} \quad \forall k \geq 0.
\]

From here, when \( \epsilon > 0 \), we have

\[
2(\theta(0) - \theta(p^*)) e^{-k\sqrt{\kappa}} \leq \epsilon \iff k \geq \sqrt{\frac{L}{\kappa} \ln \left( \frac{2(\theta(0) - \theta(p^*))}{\epsilon} \right)},
\]

while

\[
2\sqrt{L(\theta(0) - \theta(p^*))} e^{-k\sqrt{\kappa}} \leq \epsilon \iff k \geq 2 \sqrt{\frac{L}{\kappa} \ln \left( \frac{2\sqrt{L(\theta(0) - \theta(p^*))}}{\epsilon} \right)}.
\]

In conclusion, in order to guarantee \( \epsilon \)-accuracy for the dual objective values and for the decrease of \( \|\nabla \theta(\cdot)\| \) to 0, we need \( O(\ln(\frac{1}{\epsilon})) \) iterations.

6. Two examples in image processing

In this section, we solve a linear inverse problem which arises in the field of signal and image processing via the DS algorithm developed in this article. For a given matrix \( A \in \mathbb{R}^{n \times n} \) describing a blur operator and a given vector \( b \in \mathbb{R}^n \) representing the blurred and noisy image the task is to estimate the unknown original image \( x^* \in \mathbb{R}^n \) fulfilling

\[ Ax = b. \]

To this end we make use of two regularization functionals with different properties.

6.1. An \( l_1 \) regularization problem

We start by solving the \( l_1 \) regularized convex optimization problem

\[
(P) \quad \inf_{x \in S} \left\{ \|Ax - b\|^2 + \lambda \|x\|_1 \right\},
\]

where \( S \subseteq \mathbb{R}^n \) is an \( n \)-dimensional cube representing the range of the pixels and \( \lambda > 0 \) the regularization parameter. The problem to be solved can be equivalently written as

\[
(P) \quad \inf_{x \in \mathbb{R}^n} \left\{ f(x) + g(Ax) \right\},
\]

for \( f: \mathbb{R}^n \to \mathbb{R}, f(x) = \lambda \|x\|_1 + \delta_S(x) \) and \( g: \mathbb{R}^n \to \mathbb{R}, g(y) = \|y - b\|^2 \). Thus \( f \) is proper, convex and lower semicontinuous with bounded domain and \( g \) is a 2-strongly convex function with full domain, differentiable everywhere and with Lipschitz continuous...
gradient having as Lipschitz constant 2. This means that we are in the setting of Section 5.2.

By making use of gradient methods, both the iterative shrinkage-thresholding algorithm (ISTA) [9] and its accelerated variant fast ISTA FISTA [2, 3] solve the optimization problem \((P)\) in \(O(\frac{1}{C_1})\) and \(O(\frac{1}{C_0})\) iterations, respectively, whereas the convergence rate of our method is \(O\left(\frac{1}{C_0} \ln\left(\frac{1}{C_1}\right)\right)\).

Since each pixel furnishes a grey scale value which is between 0 and 255, a natural choice for the convex set \(S\) would be the \(n\)-dimensional cube \([0, 255]^n \subseteq \mathbb{R}^n\). In order to reduce the Lipschitz constant which appears in the developed approach, we scale the picture, which refer within this section, such that each of their pixel ranges in the interval \([0, \frac{1}{10}]\). We concretely look at the \(256 \times 256\) cameraman test image, which is part of the image processing toolbox in Matlab. The dimension of the vectorized and scaled cameraman test image is \(n = 256^2 = 65,536\). By making use of the Matlab functions \texttt{imfilter} and \texttt{fspecial}, this image is blurred as follows:

\begin{verbatim}
H=fspecial('gaussian',9,4); % gaussian blur of size 9 times 9
B=imfilter(X,H,'conv','symmetric'); % B=observed blurred image
\end{verbatim}

In row 1, the function \texttt{fspecial} returns a rotationally symmetric Gaussian lowpass filter of size \(9 \times 9\) with standard deviation 4. The entries of \(H\) are nonnegative and their sum adds up to 1. In row 3, the function \texttt{imfilter} convolves the filter \(H\) with the image \(X \in \mathbb{R}^{256 \times 256}\) and outputs the blurred image \(B \in \mathbb{R}^{256 \times 256}\). The boundary option “symmetric” corresponds to reflexive boundary conditions.

Thanks to the rotationally symmetric filter \(H\), the linear operator \(A \in \mathbb{R}^{256 \times 256}\) given by the Matlab function \texttt{imfilter} is symmetric too. By making use of the real spectral decomposition of \(A\), it shows that \(\|A\|^2 = 1\). After adding a zero-mean white Gaussian noise with standard deviation \(10^{-4}\), we obtain the blurred and noisy image \(b \in \mathbb{R}^n\) which is shown in Figure 1.

![Figure 1. The 256 × 256 cameraman test image: (a) original and (b) blurred and noisy.](image-url)
The dual optimization problem in minimization form is

\[
(D) \quad - \inf_{\rho \in \mathbb{R}^n} \left\{ f^*(A^* p) + g^*(-p) \right\}
\]

and, due to the fact that \( g \) has full domain, strong duality for (P) and (D) holds, i.e. \( \nu(P) = \nu(D) \) and (D) has an optimal solution (see, e.g. [5,6]). By taking into consideration (36), the smoothing parameter is taken as

\[
\rho := \frac{\epsilon}{2D_f}, \tag{38}
\]

for \( D_f = \sup \left\{ \frac{\|x\|^2}{2} : x \in \left[ 0, \frac{1}{10} \right]^n \right\} = 327.68 \), while the accuracy is chosen to be \( \epsilon = 0.3 \) and the regularization parameter is set to \( \lambda = 2e - 6 \).

We show next that the sequences of approximate primal solutions \((x_{f,p_k})_{k \geq 0}\) and \((x_{g,p_k})_{k \geq 0}\) can be easily calculated. Indeed, for \( k \geq 0 \) we have

\[
x_{f,p_k} = \arg \min_{x \in [0,1]^n} \left\{ \lambda \|x\|_1 + \frac{\rho}{2} \left( \frac{A^* p_k}{\rho} - x \right)^2 \right\}
\]

and, in order to determine it, we need to solve the one-dimensional convex optimization problem

\[
\inf_{x \in [0,1]} \left\{ \lambda x_i + \frac{\rho}{2} \left( \frac{(A^* p_k)_i}{\rho} - x_i \right)^2 \right\},
\]

for \( i = 1, \ldots, n \), which has as unique optimal solution \( \mathcal{P}_{[0,1]} \left( \frac{1}{\rho} (A^* p_k)_i - \lambda \right) \). Thus,

\[
x_{f,p_k} = \mathcal{P}_{[0,1]} \left( \frac{1}{\rho} (A^* p_k - \lambda 1^n) \right).
\]

On the other hand, for all \( k \geq 0 \) we have

\[
x_{g,p_k} = \arg \min_{x \in \mathbb{R}^n} \left\{ (p_k, x) + g(x) \right\} = \arg \min_{x \in \mathbb{R}^n} \left\{ (p_k, x) + \|x - b\|^2 \right\} = b - \frac{1}{2} p_k.
\]

Figure 2 shows the iterations 50 and 100 of ISTA, FISTA and the DS approach. The objective function values at iteration \( k \) are denoted by ISTA\(_k\), FISTA\(_k\) and, respectively, DS\(_k\) (e.g. DS\(_k := f(x_{f,p_k}) + g(A x_{f,p_k})\)). All in all, the visual quality of the restored cameraman image after 100 iterations, when using FISTA or DS, is quite comparable, whereas the recovered image by ISTA is still blurry. However, a valuable tool for measuring the quality of these images is the so-called improvement in signal-to-noise ratio (ISNR), which is defined as

\[
\text{ISNR}(k) = 10 \log_{10} \left( \frac{\|x - b\|^2}{\|x - x_k\|^2} \right),
\]
Figure 2. Iterations of ISTA, FISTA and DS for solving (P) when applied to the cameraman test image file.
where $x$, $b$ and $x_k$ denote the original, observed and estimated image at iteration $k$, respectively. Figure 3 shows the evolution of the ISNR values when using DS, FISTA and ISTA to solve $(P)$.

### 6.2. An $l_2-l_1$ regularization problem

The second convex optimization problem we solve is

$$(P) \quad \inf_{x \in S} \{ \|Ax - b\|^2 + \lambda(\|x\|^2 + \|x\|_1) \},$$

where $S \subseteq \mathbb{R}^n$ is the $n$-dimensional cube $[0, 1]^n$ representing the pixel range, $\lambda > 0$ the regularization parameter and $\|\cdot\|^2 + \|\cdot\|_1$ the regularization functional, already used in [7]. The problem to be solved can be equivalently written as

$$(P) \quad \inf_{x \in \mathbb{R}^n} \{ f(x) + g(Ax) \},$$

for $f: \mathbb{R}^n \to \mathbb{R}$, $f(x) = \lambda(\|x\|^2 + \|x\|_1) + \delta_S(x)$ and $g: \mathbb{R}^n \to \mathbb{R}$, $g(y) = \|y - b\|^2$. Thus $f$ is proper, $2\lambda$-strongly convex and lower semicontinuous with bounded domain and $g$ is a 2-strongly convex function with full domain, differentiable everywhere and with Lipschitz continuous gradient having as Lipschitz constant 2. This time we are in the setting of the Section 5.3, the Lipschitz constant of the gradient of $\theta: \mathbb{R}^n \to \mathbb{R}$, $\theta(p) = f^*(A^*p) + g^*(-p)$, being $L = \frac{1}{2\lambda} + \frac{1}{2}$. By applying the DS approach one obtains a rate of convergence of $O(\ln(\frac{1}{\epsilon}))$ for solving $(P)$.

In this example, we take a look at the blobs test image shown in Figure 4 which is also part of the image processing toolbox in Matlab. The picture undergoes the same blur as described in the previous section. Since our pixel range has changed, we now...
use additive zero-mean white Gaussian noise with standard deviation $10^{-3}$ and the regularization parameter is changed to $\lambda = 2e - 5$.

We calculate next the sequences of approximate primal solutions $(x_{f,pk})_{k \geq 0}$ and $(x_{g,pk})_{k \geq 0}$. Indeed, for $k \geq 0$ we have

$$x_{f,pk} = \arg \min_{x \in [0,1]^n} \left\{ \lambda \|x\|^2 + \lambda \|x\|_1 - \langle A^*p_k, x \rangle \right\}$$
$$= \arg \min_{x \in [0,1]^n} \left\{ \sum_{i=1}^n \left[ -(A^*p_k)_i x_i + \lambda x_i^2 + \lambda x_i \right] \right\} = \mathcal{P}_{[0,1]^n} \left( \frac{1}{2\lambda} (A^*p_k - \lambda 1^n) \right)$$

and

$$x_{g,pk} = \arg \min_{x \in \mathbb{R}^n} \{ p_k + g(x) \} = \arg \min_{x \in \mathbb{R}^n} \{ \|p_k, x\| + \|x - b\|^2 \} = b - \frac{1}{2} p_k.$$  

Figure 5 shows the iterations 50 and 100 of ISTA, FISTA and the DS technique together with the corresponding function values denoted by ISTA$_{k}$, FISTA$_{k}$ or DS$_{k}$. As before, the function values of FISTA are slightly lower than those of DS, while ISTA is far behind these methods, not only from theoretical point of view, but also can be detected visually. Figure 6 displays the ISNR for ISTA, FISTA and DS and it shows that DS outperforms the other two methods from the point of view of the quality of the reconstruction.

7. Conclusion

In this article, we have investigated the possibilities of accelerating the DS technique when solving unconstrained nondifferentiable convex optimization problems. This method, which assumes the minimization of the doubly regularized Fenchel dual objective, allows in the most general case to reconstruct an approximately optimal primal solution in $O\left(\frac{1}{\epsilon} \ln \left(\frac{1}{\epsilon}\right)\right)$ iterations. We show that under some appropriate assumptions for the functions involved in the formulation of the problem to be solved, this convergence rate can be improved to $O\left(\frac{1}{\sqrt{\epsilon}} \ln \left(\frac{1}{\epsilon}\right)\right)$, or even to $O(\ln \left(\frac{1}{\epsilon}\right))$.  

Figure 4. The $272 \times 329$ blobs test image: (a) original and (b) blurred and noisy.
Figure 5. Iterations of ISTA, FISTA and DS for solving (P) when applied to the blobs test image file.
Acknowledgements
This research was partially supported by DFG (German Research Foundation), project BO 2516/4-1 and Graduate Fellowship of the Free State Saxony, Germany.

References

[1] H.H. Bauschke and P.L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, CMS Books in Mathematics, Springer, New York, 2011.
[2] A. Beck and M. Teboulle, A fast iterative shrinkage-tresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009), pp. 183–202.
[3] A. Beck and M. Teboulle, Gradient-based algorithms with applications to signal recovery problems, in Convex Optimization in Signal Processing and Communications, Y. Eldar and D. Palomar, eds., Cambridge University Press, Cambridge, 2010, pp. 33–88.
[4] J.F. Bonnans and A. Shapiro, Perturbation Analysis of Optimization Problems, Springer Series in Operations Research and Financial Engineering, Springer, New York, 2000.
[5] R.I. Boţ, Conjugate Duality in Convex Optimization, Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer-Verlag, Berlin, Heidelberg, 2010.
[6] R.I. Boţ, S.-M. Grad, and G. Wanka, Duality in Vector Optimization, Springer-Verlag, Berlin, Heidelberg, 2009.
[7] R.I. Boţ and T. Hein, Iterative regularization with general penalty term theory and application to $L^1$- and TV-regularization, Inverse Probl. 28 (2012), p. 19, 10410.
[8] R.I. Boţ and C. Hendrich, A double smoothing technique for solving unconstrained nondifferentiable convex optimization problems (2012). Available at arXiv:1203.2070v1 [math.OC].
[9] I. Daubechies, M. Defrise, and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Commun. Pure Appl. Math. 57 (2004), pp. 1413–1457.
[10] O. Devolder, F. Glineur, and Y. Nesterov, *Double smoothing technique for large-scale linearly constrained convex optimization*, SIAM J. Optim. 22 (2012), pp. 702–727.

[11] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic, Dordrecht, 2004.

[12] Y. Nesterov, *Excessive gap technique in nonsmooth convex optimization*, SIAM J. Optim. 16 (2005), pp. 235–249.

[13] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program. 103 (2005), pp. 127–152.

[14] Y. Nesterov, *Smoothing technique and its applications in semidefinite optimization*, Math. Program. 110 (2005), pp. 245–259.