A general double-proximal gradient algorithm for d.c. programming

Sebastian Banert [*] Radu Ioan Boț [†]

October 20, 2016

The possibilities of exploiting the special structure of d.c. programs, which consist of optimizing the difference of convex functions, are currently more or less limited to variants of the DCA proposed by Pham Dinh Tao and Le Thi Hoai An in 1997. These assume that either the convex or the concave part, or both, are evaluated by one of their subgradients.

In this paper we propose an algorithm which allows the evaluation of both the concave and the convex part by their proximal points. Additionally, we allow a smooth part, which is evaluated via its gradient. In the spirit of primal-dual splitting algorithms, the concave part might be the composition of a concave function with a linear operator, which are, however, evaluated separately.

For this algorithm we show that every cluster point is a solution of the optimization problem. Furthermore, we show the connection to the Toland dual problem and prove a descent property for the objective function values of a primal-dual formulation of the problem. Convergence of the iterates is shown if this objective function satisfies the Kurdyka–Łojasiewicz property. In the last part, we apply the algorithm to an image processing model.

Key Words. d.c. programming, Toland dual, proximal-gradient algorithm, Kurdyka–Łojasiewicz property, convergence analysis

AMS subject classification. 90C26, 90C30, 65K05

[*] University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: sebastian.banert@univie.ac.at. Research supported by FWF (Austrian Science Fund), project I 2419-N32.

[†] University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: radu.bot@univie.ac.at. Research supported by FWF (Austrian Science Fund), project I 2419-N32.

1 Introduction

Optimization problems where the objective function can be written as a difference of two convex functions arise naturally in several applications, such as image processing [17], machine learning [24], optimal transport [10] and sparse signal recovering [13]. Generally, the class of d.c. functions is rather broad and contains for example every twice continuously differentiable function. For an overview over d.c. functions, see e.g. [14].

The classical approach to iteratively find local extrema of d.c. problems was described by Tao and An [23] in 1997 under the name DCA (d.c. algorithms). One of the most recent papers on this topic is [2], where an accelerated variant of the DCA method is proposed under the supplementary assumption that both the convex and the concave part are continuously differentiable. In 2003, Sun, Sampaio and Candido introduced a proximal point approach into the theory of d.c. algorithm [22], where the convex part is evaluated by its proximal point operator, while its concave part is still evaluated by one of its subgradients. Later on, the approach in [22] has been extended in [18, 1, 11] by considering in the convex part a further convex smooth summand that is evaluated via its gradient.

In this paper, we go one step further by proposing an algorithm, where both convex and concave part are evaluated via proximal steps. In convex optimisation, using proximal steps instead of subgradient steps has several advantages:

- The subdifferential at a point may be a non-singleton set, in particular it may be empty or may consist of several distinct elements. In an algorithm, one may get stuck or have to choose one, respectively.
- Even if the subdifferential is a singleton in each step, it might be highly discontinuous, so small deviations might lead to a very different behaviour of the iterations.
- Better convergence rates can be guaranteed for proximal algorithms than for subgradient algorithms (compare [6] and [19, Theorem 3.2.3]).

In addition, we consider a linear operator in the concave part of the objective function, which is evaluated in a forward manner in the spirit of primal-dual splitting methods.

In Section 2, we present the problem to be solved together with its Toland dual and attach to them a primal-dual formulation in form of a minimization problem, too. We derive first-order optimality conditions and relate the optimal solutions and the critical points of the primal-dual minimization problems to the optimal solutions and, respectively, the critical points of both primal and dual optimization problems.

In Section 3, we propose a double-proximal d.c. algorithm, which generates both a primal and a dual sequence of iterates and show several properties which make it comparable to DCA. More precisely, we prove a descent property for the objective function values of a primal-dual formulation and that every cluster point of the sequence of primal iterates is a critical point of the primal problem, while every critical point of the sequence of dual iterates is a critical point of the dual problem.

In Section 4, we show global convergence of our algorithm and convergence rates for the iterates in some certain cases, provided that the objective function of the primal-dual reformulation satisfies the Kurdyka–Łojasiewicz property; in other words, it is a *KŁ function.* The convergence analysis relies on methods and concepts of real algebraic geometry introduced by Łojasiewicz [16] and Kurdyka [15] and later developed in the nonsmooth setting by Attouch, Bolte and Svaiter [4] and Bolte, Sabach and Teboulle [8]. One of the remarkable properties of the KŁ functions is their ubiquity in applications (see [8]). The class of KŁ functions contains semi-algebraic, real sub-analytic, semiconvex, uniformly convex and convex functions satisfying a growth condition.

We close our paper with some numerical examples addressing an image deblurring and denoising problem in the context of different DC regularizations.

1.1 Notation and preliminaries

For the theory of convex analysis in finite-dimensional spaces, see the book [20]. We shall consider functions taking values in the *extended real line* $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$. We agree on the order $-\infty < a < +\infty$ for any real number a and the operations

$$\begin{aligned} +\infty + a = a + \infty = +\infty - \infty = -\infty + \infty = +\infty + \infty = +\infty, \\ -\infty + a = a - \infty = -\infty - \infty = -\infty, \\ 0 \cdot (-\infty) = 0, 0 \cdot (+\infty) = +\infty \end{aligned}$$

for arbitrary $a \in \mathbb{R}$ (see [28]). Let $\mathcal{H}$ be a real finite-dimensional Hilbert space. For a function $f : \mathcal{H} \to \overline{\mathbb{R}}$, we denote by

$$\operatorname{dom} f := \{x \in \mathcal{H} \,|\, f(x) < +\infty\}$$

its *domain.* The function f is called *proper* if it does not take the value $-\infty$ and $\operatorname{dom} f \neq \emptyset$. It is called *convex* if

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

for all $x, y \in \mathcal{H}$ and $0 \leq \lambda \leq 1$. The *conjugate function* $f^* : \mathcal{H} \to \overline{\mathbb{R}}$ of $f : \mathcal{H} \to \overline{\mathbb{R}}$ is defined by

$$f^*(x^*) = \sup \{\langle x^*, x\rangle - f(x) \,|\, x \in \mathcal{H}\}.$$

If f is proper, convex and lower semicontinuous, then $f^{**} := (f^*)^* = f$ by the Fenchel–Moreau theorem.

The *convex subdifferential* $\partial f(x)$ at $x \in \mathcal{H}$ of a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is empty if $x \notin \operatorname{dom} f$ and

$$\partial f(x) = \{x^* \in \mathcal{H} \,|\, \forall y \in \mathcal{H} : f(y) \geq f(x) + \langle x^*, y - x\rangle\}$$

otherwise. Let $\gamma > 0$ and $f : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. The *proximal point* $\operatorname{Prox}_{\gamma f}(x)$ of γf at $x \in \mathcal{H}$ is defined as

$$\operatorname{Prox}_{\gamma f}(x) = \arg\min \left\{ \gamma f(y) + \frac{1}{2}\|y - x\|^2 \,\middle|\, y \in \mathcal{H} \right\}.$$

The set of minimizers in the definition above is a singleton [5, Proposition 12.15], and the proximal point is characterised by the variational inequality [5, Proposition 12.26]

$$f(y) \geq f(\operatorname{Prox}_{\gamma f}(x)) + \frac{1}{2\gamma} \langle y - \operatorname{Prox}_{\gamma f}(x), x - \operatorname{Prox}_{\gamma f}(x) \rangle$$

for all $y \in \mathcal{H}$, which is equivalent to

$$\frac{1}{\gamma}(x - \operatorname{Prox}_{\gamma f}(x)) \in \partial f(\operatorname{Prox}_{\gamma f}(x)). \tag{1}$$

When dealing with nonconvex and nonsmooth functions, we have to consider subdifferentials more general than the convex one. The *Fréchet subdifferential* $\partial_F f(x)$ at $x \in \mathcal{H}$ of a proper and lower semicontinuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is empty if $x \notin \operatorname{dom} f$ and

$$\partial_F f(x) = \left\{ x^* \in \mathcal{H} \,\middle|\, \liminf_{\substack{y \to x \\ y \neq x}} \frac{f(y) - f(x) - \langle x^*, y - x \rangle}{\|y - x\|} \geq 0 \right\}$$

otherwise. The *limiting (Mordukhovich) subdifferential* $\partial_L f(x)$ at $x \in \mathcal{H}$ of a proper and lower semicontinuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is empty if $x \notin \operatorname{dom} f$ and

$$\partial_L f(x) = \Big\{ x^* \in \mathcal{H} \,\Big|\, \exists (x_k)_{k \geq 0}, (x_k^*)_{k \geq 0} : x_k \in \mathcal{H}, x_k^* \in \partial_F f(x_k), k \geq 0, \\ x_k \to x, f(x_k) \to f(x), x_k^* \to x^* \text{ as } k \to +\infty \Big\}$$

otherwise.

2 Problem statement

Let $\mathcal{G}$ and $\mathcal{H}$ be real finite-dimensional Hilbert spaces, let $g : \mathcal{H} \to \overline{\mathbb{R}}$ and $h : \mathcal{G} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous functions, let $\varphi : \mathcal{H} \to \mathbb{R}$ be a convex, Fréchet differentiable function with $\frac{1}{\beta}$-Lipschitz continuous gradient, for some $\beta > 0$, and let $K : \mathcal{H} \to \mathcal{G}$ be a linear mapping (and $K^* : \mathcal{G} \to \mathcal{H}$ its adjoint). We consider the problem

$$\min \{ g(x) + \varphi(x) - h(Kx) \,|\, x \in \mathcal{H} \} \tag{2}$$

together with its Toland dual problem [26, 25]

$$\min \{ h^*(y) - (g + \varphi)^*(K^* y) \,|\, y \in \mathcal{G} \}. \tag{3}$$

The following primal-dual formulation will turn out to be useful in the sequel:

$$\min \{ \Phi(x, y) \,|\, x \in \mathcal{H}, y \in \mathcal{G} \} \qquad \text{with } \Phi(x, y) := g(x) + \varphi(x) + h^*(y) - \langle y, Kx \rangle, \tag{4}$$

where $\Phi : \mathcal{H} \times \mathcal{G} \to \overline{\mathbb{R}}$ is proper and lower semicontinuous.

Let us derive necessary optimality conditions for the problems (2), (3) and (4):

Proposition 1. 1. *The optimal values of* (2), (3) *and* (4) *are equal.*

2. *For all $x \in \mathcal{H}$ and $y \in \mathcal{G}$,*

$$\begin{aligned}\Phi(x,y) &\geq g(x) + \varphi(x) - h(Kx) \qquad \textit{and}\\ \Phi(x,y) &\geq h^*(y) - (g+\varphi)^*(y)(K^*y).\end{aligned}$$

3. *Let $\bar{x} \in \mathcal{H}$ be a solution of* (2). *Then $\partial(h \circ K)(\bar{x}) \subseteq \partial g(\bar{x}) + \nabla\varphi(\bar{x})$.*

4. *Let $\bar{y} \in \mathcal{G}$ be a solution of* (3). *Then $\partial((g+\varphi)^* \circ K^*)(\bar{y}) \subseteq \partial h^*(\bar{y})$.*

5. *Let $(\bar{x}, \bar{y}) \in \mathcal{H} \times \mathcal{G}$ be a solution of* (4). *Then $\bar{x}$ is a solution of* (2), *and $\bar{y}$ is a solution of* (3). *Furthermore, the inclusions*

$$K^*\bar{y} \in \partial g(\bar{x}) + \nabla\varphi(\bar{x}), \tag{5}$$

$$K\bar{x} \in \partial h^*(\bar{y}) \tag{6}$$

hold.

Proof. 1. By the Fenchel–Moreau theorem, applied to h, we have

$$\begin{aligned}&\inf\{g(x) + \varphi(x) - h(Kx) \,|\, x \in \mathcal{H}\}\\ &= \inf\{g(x) + \varphi(x) - h^{**}(Kx) \,|\, x \in \mathcal{H}\}\\ &= \inf\{g(x) + \varphi(x) - \sup\{\langle y, Kx\rangle - h^*(y) \,|\, y \in \mathcal{G}\} \,|\, x \in \mathcal{H}\}\\ &= \inf\{g(x) + \varphi(x) + h^*(y) - \langle y, Kx\rangle \,|\, x \in \mathcal{H}, y \in \mathcal{G}\}\\ &= \inf\{h^*(y) - \sup\{\langle x, K^*y\rangle - (g+\varphi)(x) \,|\, x \in \mathcal{H}\} \,|\, y \in \mathcal{G}\}\\ &= \inf\{h^*(y) - (g+\varphi)^*(K^*y) \,|\, y \in \mathcal{G}\}.\end{aligned}$$

2. Let $x \in \mathcal{H}$ and $y \in \mathcal{G}$. Then,

$$\begin{aligned}g(x) + \varphi(x) - h(Kx) &= g(x) + \varphi(x) - h^{**}(Kx)\\ &= g(x) + \varphi(x) - \sup\{\langle Kx, \tilde{y}\rangle - h^*(\tilde{y}) \,|\, \tilde{y} \in \mathcal{G}\}\\ &\leq g(x) + \varphi(x) - \langle Kx, y\rangle + h^*(y),\end{aligned}$$

and the other inequality is verified by an analogous calculation.

3. Let $\bar{x} \in \mathcal{H}$ be a solution of (2), i.e.,

$$\forall x \in \mathcal{H} : g(x) + \varphi(x) - h(Kx) \geq g(\bar{x}) + \varphi(\bar{x}) - h(K\bar{x}). \tag{7}$$

If $h(K\bar{x}) = +\infty$, then, by definition, $\partial(h \circ K)(\bar{x}) = \emptyset$, and the inclusion automatically holds. If $h(K\bar{x}) < +\infty$, then the optimal value of (2) must be $> -\infty$, which implies

$$h(Kx) < +\infty \qquad \text{for all } x \in \operatorname{dom} g. \tag{8}$$

Now let $y \in \partial(h \circ K)(\bar{x})$. Then

$$\forall x \in \mathcal{H} : h(Kx) \geq h(K\bar{x}) + \langle y, x - \bar{x} \rangle . \tag{9}$$

Adding (7) and (9) yields

$$\forall x \in \mathcal{H} : g(x) + \varphi(x) \geq g(\bar{x}) + \varphi(\bar{x}) + \langle y, x - \bar{x} \rangle . \tag{10}$$

If $g(x) = +\infty$, then (10) is automatically satisfied, otherwise $x \in \operatorname{dom} g$ and, by (8), $h(Kx) < +\infty$, and both sides of both (7) and (9) are finite. In either case, we have shown $y \in \partial(g + \varphi)(\bar{x}) = \partial g(\bar{x}) + \nabla\varphi(\bar{x})$.

4. The proof of this statement is analogous.

5. Let $(\bar{x}, \bar{y})$ be a solution of (4). (In particular, if such a solution exists, the common optimal value of (2), (3) and (4) must be finite.) The function $x \mapsto \Phi(x, \bar{y})$ is convex and takes a minimum at $\bar{x}$. Therefore

$$0 \in \partial g(\bar{x}) + \nabla\varphi(\bar{x}) - K^*\bar{y},$$

which proves (5). The same argument works for the function $y \mapsto \Phi(\bar{x}, y)$ and implies

$$0 \in \partial h^*(\bar{y}) - K\bar{x},$$

which is (6). For these inclusions, we obtain equality in the Young–Fenchel inequality, i.e.,

$$\begin{aligned} (g + \varphi)^*(K^*\bar{y}) + (g + \varphi)(\bar{x}) &= \langle \bar{x}, K^*\bar{y} \rangle , \\ h^*(\bar{y}) + h(K\bar{x}) &= \langle \bar{y}, K\bar{x} \rangle . \end{aligned}$$

Therefore,

$$\begin{aligned} (g + \varphi)(\bar{x}) - h(K\bar{x}) &= h^*(\bar{y}) - (g + \varphi)^*(K^*\bar{y}) \\ &= h^*(\bar{y}) - \sup \{ \langle x, K^*\bar{y} \rangle - g(x) - \varphi(x) \,|\, x \in \mathcal{H} \} \\ &\leq h^*(\bar{y}) + g(\bar{x}) + \varphi(\bar{x}) - \langle \bar{x}, K^*\bar{y} \rangle . \end{aligned}$$

Since $(\bar{x}, \bar{y})$ is a solution of (4), the last expression equals the common optimal value of (2), (3) and (4). □

Definition 1. We say that $(\bar{x}, \bar{y}) \in \mathcal{H} \times \mathcal{G}$ is a *critical point* of the objective function Φ of (4) if the inclusions (5) and (6) are satisfied. We denote by $\operatorname{crit}\Phi$ the set of critical points of the function Φ.

Remark 1. If $(\bar{x}, \bar{y}) \in \mathcal{H} \times \mathcal{G}$ is a critical point of Φ, then

$$K^*\bar{y} \in K^*\partial h(K\bar{x}) \cap (\partial g(\bar{x}) + \nabla\varphi(\bar{x})), \tag{11}$$

$$K\bar{x} \in K\partial(g + \varphi)^*(K^*\bar{y}) \cap \partial h^*(\bar{y}). \tag{12}$$

By adopting the terminology of e.g. [23, p. 297], we denote by

$$\operatorname{crit}(g+\varphi-h\circ K):=\{x\in\mathcal{H}:K^*\partial h(Kx)\cap(\partial g(x)+\nabla\varphi(x))\neq\emptyset\}$$

the set of critical points of the objective function $g+\varphi-h\circ K$ of (2) and by

$$\operatorname{crit}(h^*-(g+\varphi)^*\circ K^*):=\{y\in\mathcal{G}:K\partial(g+\varphi)^*(K^*y)\cap\partial h^*(y)\neq\emptyset\}$$

the set of critical points of the objective function $h^*-(g+\varphi)^*\circ K^*$ of (3). (Recall that $K^*\partial h(Kx)\subseteq\partial(h\circ K)(x)$ and $K\partial(g+\varphi)^*(K^*y)\subseteq\partial((g+\varphi)\circ K^*)(y)$.)

Thus, if $(\bar{x},\bar{y})\in\mathcal{H}\times\mathcal{G}$ is a critical point of the objective function Φ, then $\bar{x}$ is a critical point of $g+\varphi-h\circ K$ and $\bar{y}$ is a critical point of $h^*-(g+\varphi)^*\circ K^*$.

3 The algorithm

Let $(x_0,y_0)\in\mathcal{H}\times\mathcal{G}$, and let $(\gamma_n)_{n\geq 0}$ and $(\mu_n)_{n\geq 0}$ be sequences of positive numbers. We propose the following iterative scheme: For all $n\geq 0$ set

$$x_{n+1}=\operatorname{Prox}_{\gamma_n g}(x_n+\gamma_n K^*y_n-\gamma_n\nabla\varphi(x_n)), \tag{13}$$

$$y_{n+1}=\operatorname{Prox}_{\mu_n h^*}(y_n+\mu_n Kx_{n+1}). \tag{14}$$

By the inequalities for the proximal points, we have, for every $x,y\in\mathcal{H}$ and $n\geq 0$,

$$\begin{aligned}
g(x_{n+1})-g(x)&\leq\frac{1}{\gamma_n}\left\langle x_n+\gamma_n K^*y_n-\gamma_n\nabla\varphi(x_n)-x_{n+1},x_{n+1}-x\right\rangle\\
&=\frac{1}{\gamma_n}\left\langle x_n-x_{n+1},x_{n+1}-x\right\rangle+\left\langle K^*y_n,x_{n+1}-x\right\rangle-\left\langle\nabla\varphi(x_n),x_{n+1}-x\right\rangle,\\
h^*(y_{n+1})-h^*(y)&\leq\frac{1}{\mu_n}\left\langle y_n+\mu_n Kx_{n+1}-y_{n+1},y_{n+1}-y\right\rangle\\
&=\frac{1}{\mu_n}\left\langle y_n-y_{n+1},y_{n+1}-y\right\rangle+\left\langle Kx_{n+1},y_{n+1}-y\right\rangle.
\end{aligned}$$

Moreover, using [5, Theorem 18.15 (iii)] and the subdifferential inequality, we have for every $x\in\mathcal{H}$ and $n\geq 0$,

$$\begin{aligned}
\varphi(x_{n+1})-\varphi(x_n)&\leq\left\langle\nabla\varphi(x_n),x_{n+1}-x_n\right\rangle+\frac{1}{2\beta}\left\|x_n-x_{n+1}\right\|^2,\\
\varphi(x_n)-\varphi(x)&\leq\left\langle\nabla\varphi(x_n),x_n-x\right\rangle.
\end{aligned}$$

We consider the auxiliary function $\Phi:\mathcal{H}\times\mathcal{G}\to\overline{\mathbb{R}}$ defined by

$$\Phi(x,y)=g(x)+\varphi(x)+h^*(y)-\langle y,Kx\rangle.$$

By the inequalities above, we have, for arbitrary $x\in\mathcal{H}$, $y\in\mathcal{G}$ and $n\geq 0$,

$$\begin{aligned}
&\Phi(x_{n+1},y_{n+1})-\Phi(x,y)\\
&=g(x_{n+1})-g(x)+\varphi(x_{n+1})-\varphi(x)+h^*(y_{n+1})-h^*(y)+\langle y,Kx\rangle-\langle y_{n+1},Kx_{n+1}\rangle\\
&\leq\frac{1}{\gamma_n}\left\langle x_n-x_{n+1},x_{n+1}-x\right\rangle+\frac{1}{\mu_n}\left\langle y_n-y_{n+1},y_{n+1}-y\right\rangle+\frac{1}{2\beta}\left\|x_n-x_{n+1}\right\|^2\\
&\quad+\left\langle K(x-x_{n+1}),y-y_n\right\rangle.
\end{aligned}\tag{15}$$

Furthermore, for any $n \geq 0$,

$$\begin{aligned}\Phi(x_{n+1}, y_n) - \Phi(x_n, y_n) &= g(x_{n+1}) + \varphi(x_{n+1}) - g(x_n) - \varphi(x_n) + \langle K^* y_n, x_n - x_{n+1}\rangle \\ &\leq \left(\frac{1}{2\beta} - \frac{1}{\gamma_n}\right) \|x_n - x_{n+1}\|^2, \qquad (16) \\ \Phi(x_{n+1}, y_{n+1}) - \Phi(x_{n+1}, y_n) &= h^*(y_{n+1}) - h^*(y_n) + \langle y_n - y_{n+1}, Kx_{n+1}\rangle \\ &\leq -\frac{1}{\mu_n} \|y_n - y_{n+1}\|^2. \qquad (17)\end{aligned}$$

The last two inequalities give rise to the following statement.

Proposition 2. *For each $n \geq 0$, we have*

$$\Phi(x_{n+1}, y_{n+1}) \leq \Phi(x_{n+1}, y_n) \leq \Phi(x_n, y_n),$$

provided that $0 < \gamma_n \leq 2\beta$.

Proposition 3. *Let*

$$0 < \inf_{n\geq 0} \gamma_n \leq \sup_{n\geq 0} \gamma_n < 2\beta \qquad \textit{and} \qquad 0 < \inf_{n\geq 0} \mu_n \leq \sup_{n\geq 0} \mu_n < +\infty. \qquad (18)$$

Furthermore, let $\inf\{g(x) + \varphi(x) - h(Kx) \mid x \in \mathcal{H}\} > -\infty$. Then,

$$\sum_{n\geq 0} \|x_n - x_{n+1}\|^2 < +\infty \qquad \textit{and} \qquad \sum_{n\geq 0} \|y_n - y_{n+1}\|^2 < +\infty.$$

Proof. Let $N \geq 1$ be an integer. Sum up (16) and (17) for $n = 0, \ldots, N-1$ and obtain

$$\Phi(x_N, y_N) - \Phi(x_0, y_0) \leq \sum_{n=0}^{N-1} \left(\frac{1}{2\beta} - \frac{1}{\gamma_n}\right) \|x_n - x_{n+1}\|^2 - \sum_{n=0}^{N-1} \frac{1}{\mu_n} \|y_n - y_{n+1}\|^2.$$

By assumption, the expression on the left-hand side is bounded below by a fixed real number M for any $N \geq 1$, and so is the right-hand side. The numbers $\left(\frac{1}{\gamma_n} - \frac{1}{2\beta}\right)$ and $\frac{1}{\mu_n}$ are bounded below by a positive number, say $\varepsilon > 0$, so

$$\sum_{n=0}^{N-1} \|x_n - x_{n+1}\|^2 + \sum_{n=0}^{N-1} \|y_n - y_{n+1}\|^2 \leq -\frac{M}{\varepsilon}.$$

Since N is arbitrary, the series converge. □

Proposition 4. *Let $\inf\{g(x) + \varphi(x) - h(Kx) \mid x \in \mathcal{H}\} > -\infty$ and (18) be satisfied. If $(x_n)_{n\geq 0}$ and $(y_n)_{n\geq 0}$ are bounded, then*

1. *every cluster point of $(x_n)_{n\geq 0}$ is a critical point of (2),*
2. *every cluster point of $(y_n)_{n\geq 0}$ is a critical point of (3) and*

3. *every cluster point of* $(x_n, y_n)_{n\geq 0}$ *is a critical point of* (4).

Proof. Let $\bar{x}$ be a cluster point of $(x_n)_{n\geq 0}$. Let $(x_{n_k})_{k\geq 0}$ be a subsequence of $(x_n)_{n\geq 0}$ such that $x_{n_k} \to \bar{x}$. By another transition to a subsequence, we can guarantee $y_{n_k} \to \bar{y}$ for some $\bar{y} \in \mathcal{H}$, since $(y_{n_k})_{k\geq 0}$ is bounded. By (13) and (14), we obtain, for every $k \geq 0$,

$$\frac{x_{n_k} - x_{n_k+1}}{\gamma_{n_k}} + K^* y_{n_k} - \nabla\varphi(x_{n_k}) \in \partial g(x_{n_k+1})$$
$$\text{and } \frac{y_{n_k} - y_{n_k+1}}{\mu_{n_k}} + K x_{n_k+1} \in \partial h^*(y_{n_k+1}),$$

respectively. By Proposition 3, the first summands on the left-hand side of the above inclusions tend to zero as $k \to \infty$. Using the continuity of $\nabla\varphi$ and the closedness of the graphs of ∂g and ∂h^* and passing to the limit, we get $K^*\bar{y} - \nabla\varphi(\bar{x}) \in \partial g(\bar{x})$ and $K\bar{x} \in \partial h^*(\bar{y})$, which means that $(\bar{x}, \bar{y})$ is a critical point of Φ. The first statement follows by considering Remark 1. For the second statement, one has to choose $\bar{x}$ and $\bar{y}$ in reverse order, for the third one, they are chosen at the same time. □

Remark 2. It is clear that one cannot expect the cluster points to be minima, since it is easy to see that $(\bar{x}, \bar{y})$ is a fixed point of the iteration (13)–(14) if and only if (5) and (6) are satisfied, i.e., if and only if $(\bar{x}, \bar{y})$ is a critical point for Φ (independent of the choice of the parameters $(\gamma_n)_{n\geq 0}$ and $(\mu_n)_{n\geq 0}$).

Proposition 5. *Let* (18) *be satisfied. For any* $n \geq 0$*, the following statements are equivalent:*

1. (x_n, y_n) *is a critical point of* Φ*;*
2. $(x_{n+1}, y_{n+1}) = (x_n, y_n)$*;*
3. $\Phi(x_{n+1}, y_{n+1}) = \Phi(x_n, y_n)$.

Proof. It is easily seen by the formula (1) that the first two items are equivalent. The equivalence of the latter two items follows by (16) and (17). □

Next, we summarise the convergence properties of the prox-prox algorithm. To this end, we denote by $\omega(x_0, y_0)$ the set of cluster points of the iteration generated by (13) and (14) with the initial points x_0 and y_0. See also [8, Lemma 5] for an analogous result for a nonconvex forward-backward scheme.

Lemma 1. *Let* $\mathcal{H}$ *and* $\mathcal{G}$ *be two real finite-dimensional Hilbert spaces, let* $g : \mathcal{H} \to \overline{\mathbb{R}}$ *and* $h : \mathcal{G} \to \overline{\mathbb{R}}$ *be proper, convex and lower semicontinuous functions, let* $\varphi : \mathcal{H} \to \mathbb{R}$ *be a convex, Fréchet differentiable function with a* $\frac{1}{\beta}$*-Lipschitz continuous gradient, for some* $\beta > 0$*, and let* $K : \mathcal{H} \to \mathcal{G}$ *be a linear mapping. Let the sequences* $(\gamma_n)_{n\geq 0}$ *and* $(\mu_n)_{n\geq 0}$ *satisfy* (18)*. Moreover, assume that the sequence* $(x_n, y_n)_{n\geq 0}$ *generated by* (13) *and* (14) *is bounded. Then the following assertions hold:*

1. $\emptyset \neq \omega(x_0, y_0) \subseteq \operatorname{crit}\Phi \subseteq \operatorname{crit}(g + \varphi - h \circ K) \times \operatorname{crit}(h^* - (g + \varphi)^* \circ K^*)$,

2. $\lim_{n\to\infty} \operatorname{dist}((x_n, y_n), \omega(x_0, y_0)) = 0$,

3. if the common optimal value of the problems (2), (3) *and* (4) *is* $> -\infty$, *then* $\omega(x_0, y_0)$ *is a nonempty, compact and connected set, and so are the sets of the limit points of the sequences* $(x_n)_{n\geq 0}$ *and* $(y_n)_{n\geq 0}$,

4. the objective function Φ *is finite and constant on* $\omega(x_0, y_0)$ *provided that the optimal value is finite.*

Proof. 1. It is clear that the set of cluster points of a bounded sequence is nonempty. That every cluster point is critical for Φ, is the statement of Proposition 4. The last inclusion is discussed in Remark 1.

2. Assume that the assertion does not hold. In this case, there exists an $\varepsilon > 0$ and a subsequence $(x_{n_k}, y_{n_k})_{k\geq 0}$ of $(x_n, y_n)_{n\geq 0}$ with $\operatorname{dist}((x_{n_k}, y_{n_k}), \omega(x_0, y_0)) > \varepsilon$ for all $k \geq 0$. The subsequence is bounded, so it has a cluster point, which is a cluster point of the original sequence $(x_n, y_n)_{n\geq 0}$ as well, thus an element of $\omega(x_0, y_0)$. This contradicts the assumption $\operatorname{dist}((x_{n_k}, y_{n_k}), \omega(x_0, y_0)) > \varepsilon$ for all $k \geq 0$.

3. Since the sequence $(x_n, y_n)_{n\geq 0}$ is bounded, the sets

$$\Omega_k := \operatorname{cl}\left(\bigcup_{n\geq k} \{(x_n, y_n)\}\right)$$

are bounded and closed, hence compact for any $k \geq 0$. Their intersection $\bigcap_{n\geq 0} \Omega_n$, which equals the set of cluster points of $(x_n, y_n)_{n\geq 0}$, is therefore compact, too. The connectedness follows from Proposition 3. See the proof of [8, Lemma 5 (iii)] for the details.

4. According to Proposition 2, the function values $\Phi(x_n, y_n)$ are monotonically decreasing, thus convergent, say $\Phi(x_n, y_n) \to \ell$. Let $(\bar{x}, \bar{y})$ be an arbitrary limit point of the sequence $(x_n, y_n)_{n\geq 0}$, and let $(x_{n_k}, y_{n_k})_{k\geq 0}$ be a subsequence converging to $(\bar{x}, \bar{y})$ as $k \to \infty$. By lower semicontinuity, we have $\Phi(\bar{x}, \bar{y}) \leq \lim_{k\to\infty} \Phi(x_{n_k}, y_{n_k}) = \ell$. On the other hand, consider (15) with $x = \bar{x}$ and $y = \bar{y}$. The right-hand side converges to 0 as we let $n \to \infty$ along the subsequence $(n_k)_{k\geq 0}$, so $\ell = \lim_{n\to\infty} \Phi(x_n, y_n) \leq \Phi(\bar{x}, \bar{y})$. □

Remark 3. To guarantee the boundedness of the iterates, one could assume that the objective function of the primal-dual minimization problem (4) is coercive, i.e., the lower level sets are bounded.

4 Convergence under Kurdyka–Łojasiewicz assumptions

In the next step, we shall assume the Kurdyka–Łojasiewicz property for the functions involved. Let us recall the definition and some basic properties. By Θ_η, for $\eta \in (0, +\infty]$, we denote the set of all concave and continuous functions $\vartheta : [0, \eta) \to \mathbb{R}$ with the following properties:

1. $\vartheta(0) = 0$,

2. ϑ is continuously differentiable on $(0, \eta)$ and continuous at 0,

3. $\vartheta'(s) > 0$ for all $s \in (0, \eta)$.

Definition 2. Let $\mathcal{H}$ be a real finite-dimensional Hilbert space, and let $\Phi : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper and lower semicontinuous function. We say that Φ satisfies the *Kurdyka–Łojasiewicz property* at $\bar{x} \in \operatorname{dom} \partial_L \Phi := \{x \in \mathcal{H} \,|\, \partial_L \Phi(x) \neq \emptyset\}$ if there exists some $\eta \in (0, +\infty]$, a neighbourhood U of $\bar{x}$ and a function $\vartheta \in \Theta_\eta$ such that for all

$$x \in U \cap \{x \in \mathcal{H} \,|\, \Phi(\bar{x}) < \Phi(x) < \Phi(\bar{x}) + \eta\}$$

the following inequality holds

$$\vartheta'(\Phi(x) - \Phi(\bar{x})) \cdot \operatorname{dist}(0, \partial_L \Phi(x)) \geq 1.$$

We call Φ a *KŁ function* if it satisfies the Kurdyka–Łojasiewicz property at each point $\bar{x} \in \operatorname{dom} \partial_L \Phi$.

The following uniform KŁ property is according to [8, Lemma 6].

Lemma 2. *Let Ω be a compact set, and let $\Phi : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper and lower semicontinuous function. Assume that Φ is constant on Ω and satisfies the KŁ property at each point of Ω. Then there exist $\varepsilon > 0$, $\eta > 0$ and $\vartheta \in \Theta_\eta$ such that for all $\bar{u} \in \Omega$ and all u in the intersection*

$$\{u \in \mathcal{H} \,|\, \operatorname{dist}(u, \Omega) < \varepsilon\} \cap \{u \in \mathcal{H} \,|\, \Phi(\bar{u}) < \Phi(u) < \Phi(\bar{u}) + \eta\} \tag{19}$$

one has

$$\vartheta'(\Phi(u) - \Phi(\bar{u})) \cdot \operatorname{dist}(0, \partial_L \Phi(u)) \geq 1.$$

In the KŁ property, we need the distance of a subgradient from zero. In our algorithm, we have the following result.

Lemma 3. *For each $n \geq 1$ with $\gamma_{n-1} < 2\beta$, there exist $(x_n^*, y_n^*) \in \mathcal{H} \times \mathcal{G}$ with $(x_n^*, y_n^*) \in \partial_L \Phi(x_n, y_n)$ and*

$$\begin{aligned}
\|x_n^*\| &\leq \|K\| \, \|y_{n-1} - y_n\| + \frac{1}{\gamma_{n-1}} \|x_{n-1} - x_n\| \, , \\
\|y_n^*\| &\leq \frac{1}{\mu_{n-1}} \|y_{n-1} - y_n\| \, .
\end{aligned}$$

Proof. From the definition of the algorithm, we have, for each $n \geq 1$,

$$\begin{aligned}
\frac{x_{n-1} - x_n}{\gamma_{n-1}} + K^* y_{n-1} - \nabla \varphi(x_{n-1}) &\in \partial g(x_n), \\
\frac{y_{n-1} - y_n}{\mu_{n-1}} + K x_n &\in \partial h^*(y_n).
\end{aligned}$$

Consider the function $\widetilde{\Phi} : \mathcal{H} \times \mathcal{G} \to \overline{\mathbb{R}}, \widetilde{\Phi}(x,y) := g(x) + \varphi(x) + h^*(y)$. By the usual calculus of the convex subdifferential and [21, Proposition 8.12], for each $n \geq 1$

$$\partial_L \widetilde{\Phi}(x_n, y_n) = (\partial g(x_n) + \nabla\varphi(x_n)) \times \partial h^*(y_n).$$

By [21, Exercise 8.8], we have for each $n \geq 1$

$$\begin{aligned}\partial_L \Phi(x_n, y_n) &= \partial_L \widetilde{\Phi}(x_n, y_n) - (K^* y_n, K x_n) \\ &= (\partial g(x_n) + \nabla\varphi(x_n) - K^* y_n) \times (\partial h^*(y_n) - K x_n),\end{aligned} \tag{20}$$

thus,

$$\begin{pmatrix} x_n^* \\ y_n^* \end{pmatrix} := \begin{pmatrix} \frac{x_{n-1}-x_n}{\gamma_{n-1}} + \nabla\varphi(x_n) - \nabla\varphi(x_{n-1}) + K^*(y_{n-1} - y_n) \\ \frac{y_{n-1}-y_n}{\mu_{n-1}} \end{pmatrix} \in \partial_L \Phi(x_n, y_n).$$

Now, we estimate for each $n \geq 1$

$$\|x_n^*\| \leq \|K\| \, \|y_{n-1} - y_n\| + \frac{1}{\gamma_{n-1}} \|(\mathrm{Id} - \gamma_{n-1}\nabla\varphi)(x_{n-1}) - (\mathrm{Id} - \gamma_{n-1}\nabla\varphi)(x_n)\| \, .$$

By the Baillon–Haddad theorem [5, Corollary 18.16], $\nabla\varphi$ is β-cocoercive. By [5, Proposition 4.33], $\mathrm{Id} - \gamma_{n-1}\nabla\varphi$ is nonexpansive for $\gamma_{n-1} < 2\beta$, which leads to the desired conclusion. □

4.1 The case when Φ is a KŁ function

Theorem 1. *Let*

$$\begin{aligned} 0 < \underline{\gamma} := \inf_{n\geq 0} \gamma_n \leq \overline{\gamma} := \sup_{n\geq 0} \gamma_n < \beta, \\ 0 < \underline{\mu} := \inf_{n\geq 0} \mu_n \leq \overline{\mu} := \sup_{n\geq 0} \mu_n < +\infty. \end{aligned}$$

Suppose that Φ is in addition a KŁ function bounded from below. Then $(x_n, y_n)_{n\geq 0}$ is a Cauchy sequence, thus convergent to a critical point of Φ.

Proof. Let $\Omega := \omega(x_0, y_0)$, and let $\ell \in \mathbb{R}$ be the value of Φ on Ω (see item 4 of Lemma 1). If $\Phi(x_n, y_n) = \ell$ for some $n \geq 0$, then, by (16) and (17), $x_{n+1} = x_n$ and $y_{n+1} = y_n$, and the assertion holds. Therefore, we assume $\Phi(x_n, y_n) > \ell$ for all $n \geq 0$.

Let $\varepsilon > 0$, $\eta > 0$ and $\vartheta \in \Theta_\eta$ be as provided by Lemma 2. Since $\Phi(x_n, y_n) \to \ell$ as $n \to +\infty$, we find $n_1 \geq 0$ with $\Phi(x_n, y_n) < \ell + \eta$ for $n \geq n_1$. Since $\operatorname{dist}((x_n, y_n), \Omega) \to 0$ as $n \to +\infty$, we find $n_2 \geq 0$ with $\operatorname{dist}((x_n, y_n), \Omega) < \varepsilon$ for $n \geq n_2$.

In the following, fix an arbitrary $n \geq n_0 := \max\{n_1, n_2, 1\}$. Then (x_n, y_n) is an element of the intersection (19). Consequently,

$$\vartheta'(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) \cdot \operatorname{dist}((0,0), \partial_L \Phi(x_n, y_n)) \geq 1. \tag{21}$$

By the concavity of ϑ, we get, for all $s \in (0, \eta)$,

$$\vartheta(s) - \vartheta(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) \leq \vartheta'(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) \cdot (s - \Phi(x_n, y_n) + \Phi(\bar{x}, \bar{y})),$$

so, setting in particular $s := \Phi(x_{n+1}, y_{n+1}) - \Phi(\bar{x}, \bar{y}) \in (0, \eta)$,

$$\begin{aligned}
&(\vartheta(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) - \vartheta(\Phi(x_{n+1}, y_{n+1}) - \Phi(\bar{x}, \bar{y}))) \cdot \|(x_n^*, y_n^*)\| \\
\geq{}& \vartheta'(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) \cdot (\Phi(x_n, y_n) - \Phi(x_{n+1}, y_{n+1})) \cdot \|(x_n^*, y_n^*)\| \\
\geq{}& \vartheta'(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y})) \cdot (\Phi(x_n, y_n) - \Phi(x_{n+1}, y_{n+1})) \cdot \operatorname{dist}((0,0), \partial_L \Phi(x_n, y_n)) \\
\geq{}& (\Phi(x_n, y_n) - \Phi(x_{n+1}, y_{n+1})).
\end{aligned}$$

Moreover, by (16) and (17),

$$\Phi(x_n, y_n) - \Phi(x_{n+1}, y_{n+1}) \geq \left(\frac{1}{\gamma_n} - \frac{1}{2\beta}\right) \|x_n - x_{n+1}\|^2 + \frac{1}{\mu_n} \|y_n - y_{n+1}\|^2 .$$

Let us define the following shorthands

$$\begin{aligned}
\delta_n &:= \sqrt{\left(\frac{1}{\gamma_n} - \frac{1}{2\beta}\right) \|x_n - x_{n+1}\|^2 + \frac{1}{\mu_n} \|y_n - y_{n+1}\|^2}, \\
\varepsilon_n &:= \vartheta(\Phi(x_n, y_n) - \Phi(\bar{x}, \bar{y}))
\end{aligned}$$

for $n \geq n_0$ to obtain the inequality

$$(\varepsilon_n - \varepsilon_{n+1}) \cdot \|(x_n^*, y_n^*)\| \geq \delta_n^2.$$

By the arithmetic-geometric inequality, for any $r > 0$ and $n \geq n_0$

$$\begin{aligned}
\delta_n &\leq \sqrt{(r \|(x_n^*, y_n^*)\|) \cdot \left(\frac{1}{r}(\varepsilon_n - \varepsilon_{n+1})\right)} \\
&\leq \frac{1}{2}\left(r \|(x_n^*, y_n^*)\| + \frac{1}{r}(\varepsilon_n - \varepsilon_{n+1})\right) \\
&\leq r \|(x_n^*, y_n^*)\| + \frac{1}{r}(\varepsilon_n - \varepsilon_{n+1}) \qquad (22)
\end{aligned}$$

(recall that, by Proposition 2 and the properties of ϑ, the sequence $(\varepsilon_n)_{n \geq n_0}$ is decreasing, so $\varepsilon_n - \varepsilon_{n+1} \geq 0$). On the other hand, by Lemma 3 and the inequality $2ab \leq a^2 + b^2$ $(a, b \geq 0)$, for any $n \geq n_0$

$$\begin{aligned}
\|(x_n^*, y_n^*)\|^2 \leq& \left(\|K\|^2 + \frac{1}{\mu_{n-1}^2}\right) \|y_{n-1} - y_n\|^2 + \frac{1}{\gamma_{n-1}^2} \|x_{n-1} - x_n\|^2 + \\
&+ \frac{2\|K\|}{\gamma_{n-1}} \|x_{n-1} - x_n\| \|y_{n-1} - y_n\| \\
\leq& \left(2\|K\|^2 + \frac{1}{\mu_{n-1}^2}\right) \|y_{n-1} - y_n\|^2 + \frac{2}{\gamma_{n-1}^2} \|x_{n-1} - x_n\|^2 \\
\leq& C_n^2 \delta_{n-1}^2, \qquad (23)
\end{aligned}$$

with

$$C_n := \sqrt{\max\left\{ \frac{\frac{2}{\gamma_{n-1}^2}}{\frac{1}{\gamma_{n-1}} - \frac{1}{2\beta}}, \frac{2\left\|K\right\|^2 + \frac{1}{\mu_{n-1}^2}}{\frac{1}{\mu_{n-1}}} \right\}}$$

$$= \sqrt{\max\left\{ \frac{4\beta}{\gamma_{n-1}(2\beta - \gamma_{n-1})}, \frac{1 + 2\left\|K\right\|^2 \mu_{n-1}^2}{\mu_{n-1}} \right\}}.$$

For all $n \geq n_0$,

$$C_n \leq C_0 := \sqrt{\max\left\{ \frac{4\beta}{\underline{\gamma}(2\beta - \overline{\gamma})}, \frac{1 + 2\left\|K\right\|^2 \overline{\mu}^2}{\underline{\mu}} \right\}}.$$

Combined with (22), we obtain

$$\delta_n \leq rC_0\delta_{n-1} + \frac{1}{r}(\varepsilon_n - \varepsilon_{n+1}). \tag{24}$$

For any $k \geq n_0 + 1$, we have, by iteration,

$$\delta_k \leq (rC_0)^{k-n_0}\delta_{n_0} + \sum_{n=0}^{k-n_0-1} \frac{(rC_0)^n}{r}(\varepsilon_{k-n} - \varepsilon_{k-n+1}),$$

therefore, for any $N \geq n_0 + 1$ and $0 < r < \frac{1}{C_0}$,

$$\begin{aligned}
\sum_{k=n_0+1}^{N} \delta_k &\leq \sum_{k=n_0+1}^{N} \left((rC_0)^{k-n_0}\delta_{n_0} + \sum_{n=0}^{k-n_0-1} \frac{(rC_0)^n}{r}(\varepsilon_{k-n} - \varepsilon_{k-n+1}) \right) \\
&= \sum_{k=0}^{N-n_0-1} (rC_0)^{k+1}\delta_{n_0} + \sum_{k=0}^{N-n_0-1} \sum_{n=0}^{k} \frac{(rC_0)^n}{r}(\varepsilon_{k+n_0-n+1} - \varepsilon_{k+n_0-n+2}) \\
&\leq \frac{rC_0\delta_{n_0}}{1 - rC_0} + \sum_{n=0}^{N-n_0-1} \frac{(rC_0)^n}{r} \sum_{k=n}^{N-n_0-1} (\varepsilon_{k+n_0-n+1} - \varepsilon_{k+n_0-n+2}) \\
&\leq \frac{rC_0\delta_{n_0}}{1 - rC_0} + \sum_{n=0}^{N-n_0-1} \frac{(rC_0)^n}{r}\varepsilon_{n_0+1} \\
&\leq \frac{rC_0\delta_{n_0}}{1 - rC_0} + \frac{\varepsilon_{n_0+1}}{r(1 - rC_0)}.
\end{aligned}$$

The right-hand side does not depend on N, thus, we conclude that $\sum_{k=n_0+1}^{\infty} \delta_k$ is finite, and so are $\sum_{k=n_0+1}^{\infty} \|x_n - x_{n+1}\|$ and $\sum_{k=n_0+1}^{\infty} \|y_n - y_{n+1}\|$. □

4.2 Convergence rates

Lemma 4. *Assume that Φ is a KŁ function with $\vartheta(t) = Mt^{1-\theta}$ for some $M > 0$ and $0 \leq \theta < 1$. Let $\bar{x}$ and $\bar{y}$ the limit points of the sequences $(x_n)_{n\geq 0}$ and $(y_n)_{n\geq 0}$, respectively (which exist due to Theorem 1). Then the following convergence rates are guaranteed:*

1. *if $\theta = 0$, then there exists $n_0 \geq 0$, such that $x_n = x_{n_0}$ and $y_n = y_{n_0}$ for $n \geq n_0$;*
2. *if $0 < \theta \leq \frac{1}{2}$, then there exist $c > 0$ and $0 \leq q < 1$ such that*
$$\|x_n - \bar{x}\| \leq cq^n \qquad \text{and} \qquad \|y_n - \bar{y}\| \leq cq^n$$
for all $n \geq 0$;
3. *if $\frac{1}{2} < \theta < 1$, then there exists $c > 0$ such that*
$$\|x_n - \bar{x}\| \leq cn^{-\frac{1-\theta}{2\theta-1}} \qquad \text{and} \qquad \|y_n - \bar{y}\| \leq cn^{-\frac{1-\theta}{2\theta-1}}$$
for all $n \geq 0$.

Proof. 1. First, let $\theta = 0$. Assume to the contrary (see Proposition 5) that for any $n \geq 0$, $(x_{n+1}, y_{n+1}) \neq (x_n, y_n)$. We have $\vartheta'(t) = M$ for all $t > 0$ and thus, by (21),
$$M \cdot \|(x_n^*, y_n^*)\| \geq 1 \text{ for any } n \geq 1,$$
which contradicts either Lemma 3 or Proposition 3.

Before considering the other cases, assume from now on that (x_n, y_n) is not a critical point of Φ for any $n \geq 0$. Notice that $\vartheta'(t) = (1-\theta)Mt^{-\theta}$. In the proof of Theorem 1, we have shown that for $0 < r < \frac{1}{C_0}$
$$\begin{aligned}\sum_{k=n_0+1}^{\infty} \delta_k &\leq \frac{rC_0\delta_{n_0}}{1-rC_0} + \frac{\varepsilon_{n_0+1}}{r(1-rC_0)} \\ &= \frac{rC_0\delta_{n_0}}{1-rC_0} + \frac{M(\Phi(x_{n_0+1}, y_{n_0+1}) - \Phi(\bar{x}, \bar{y}))^{1-\theta}}{r(1-rC_0)} \\ &= \frac{rC_0\delta_{n_0}}{1-rC_0} + \frac{M^{1+\frac{1-\theta}{\theta}}(1-\theta)^{\frac{1-\theta}{\theta}}}{r(1-rC_0)\vartheta'(\Phi(x_{n_0+1}, y_{n_0+1}) - \Phi(\bar{x}, \bar{y}))^{\frac{1-\theta}{\theta}}} \\ &\leq \frac{rC_0\delta_{n_0}}{1-rC_0} + \frac{M^{\frac{1}{\theta}}(1-\theta)^{\frac{1-\theta}{\theta}} \left\|(x_{n_0+1}^*, y_{n_0+1}^*)\right\|^{\frac{1-\theta}{\theta}}}{r(1-rC_0)},\end{aligned}$$
where the last inequality follows from the KŁ property (notice that $\Phi(x_{n_0+1}, y_{n_0+1}) - \Phi(\bar{x}, \bar{y}) > 0$ because we assumed that (x_{n_0+1}, y_{n_0+1}) is not a critical point of Φ). We can repeat this calculation for any $n \geq n_0 + 1$ instead of $n_0 + 1$, because such an n would meet the criteria according to which we chose $n_0 + 1$. Thus, we obtain from (23), for $n \geq n_0 + 1$,
$$\sum_{k=n+1}^{\infty} \delta_k \leq \frac{rC_0\delta_n}{1-rC_0} + \frac{M^{\frac{1}{\theta}}(1-\theta)^{\frac{1-\theta}{\theta}}(C_0\delta_n)^{\frac{1-\theta}{\theta}}}{r(1-rC_0)}. \tag{25}$$
The rest of the proof follows in the lines of [3, Theorem 2]:

2. Let $0 < \theta \leq \frac{1}{2}$. Then $1 \leq \frac{1-\theta}{\theta} < +\infty$, so $\delta_n \to 0$ as $n \to \infty$ implies that the first term on the right-hand side of (25) is the dominant one. Therefore, we find $n_1 \geq n_0 + 1$ and $C_1 > 0$ such that

$$\sum_{k=n+1}^{\infty} \delta_k \leq C_1 \delta_n = C_1 \left(\sum_{k=n}^{\infty} \delta_k - \sum_{k=n+1}^{\infty} \delta_k \right)$$

for any $n \geq n_1$. Thus, for any $n \geq n_1$,

$$\sum_{k=n+1}^{\infty} \delta_k \leq \frac{C_1}{1 + C_1} \sum_{k=n}^{\infty} \delta_k.$$

By induction, for any $n \geq n_1 + 1$,

$$\delta_n \leq \sum_{k=n}^{\infty} \delta_k \leq \left(\frac{C_1}{1 + C_1} \right)^{n-n_1} \sum_{k=n_1}^{\infty} \delta_k,$$

which proves the assertion.

3. Let $\frac{1}{2} < \theta < 1$. Then $0 < \frac{1-\theta}{\theta} < 1$, so $\delta_n \to 0$ as $n \to \infty$ implies that the second term on the right-hand side of (25) is the dominant one. Therefore, we find $n_1 \geq n_0 + 1$ and $C_1 > 0$ such that

$$\sum_{k=n+1}^{\infty} \delta_k \leq C_1 \delta_n^{\frac{1-\theta}{\theta}}$$

for any $n \geq n_1$. Then, for any $n \geq n_1$,

$$\left(\sum_{k=n+1}^{\infty} \delta_k \right)^{\frac{\theta}{1-\theta}} \leq C_1^{\frac{\theta}{1-\theta}} \left(\sum_{k=n}^{\infty} \delta_k - \sum_{k=n+1}^{\infty} \delta_k \right).$$

We define $h : (0, +\infty) \to \mathbb{R}$, $h(s) = s^{-\frac{\theta}{1-\theta}}$ and notice that h is monotonically decreasing as is the sequence $(\sum_{k=n}^{\infty} \delta_k)_{n \geq n_1}$. Therefore, for any $n \geq n_1$,

$$\begin{aligned}
1 &\leq C_1^{\frac{\theta}{1-\theta}} h\left(\sum_{k=n+1}^{\infty} \delta_k \right) \left(\sum_{k=n}^{\infty} \delta_k - \sum_{k=n+1}^{\infty} \delta_k \right) \\
&\leq C_1^{\frac{\theta}{1-\theta}} \int_{\sum_{k=n+1}^{\infty} \delta_k}^{\sum_{k=n}^{\infty} \delta_k} h(s) \, \mathrm{d}s \\
&= -C_1^{\frac{\theta}{1-\theta}} \frac{1-\theta}{2\theta - 1} \left(\left(\sum_{k=n}^{\infty} \delta_k \right)^{-\frac{2\theta-1}{1-\theta}} - \left(\sum_{k=n+1}^{\infty} \delta_k \right)^{-\frac{2\theta-1}{1-\theta}} \right).
\end{aligned}$$

Thus, by induction, for any $n \geq n_1 + 1$,

$$\left(\sum_{k=n}^{\infty} \delta_k\right)^{-\frac{2\theta-1}{1-\theta}} + \frac{(2\theta-1)(n-n_1)}{C_1^{\frac{\theta}{1-\theta}}(1-\theta)} \leq \left(\sum_{k=n_1}^{\infty} \delta_k\right)^{-\frac{2\theta-1}{1-\theta}}.$$

The assertion follows by

$$\delta_n \leq \sum_{k=n}^{\infty} \delta_k \leq \left(\left(\sum_{k=n_1}^{\infty} \delta_k\right)^{-\frac{2\theta-1}{1-\theta}} + \frac{(2\theta-1)(n-n_1)}{C_1(1-\theta)}\right)^{-\frac{1-\theta}{2\theta-1}} \quad \text{for any } n \geq n_1{+}1.$$

□

5 Application to image processing

Consider an image of the size $m \times n$ pixels. (For the sake of simplicity, we consider gray-scale pictures only.) It can be represented by a vector $x \in \mathcal{H} := \mathbb{R}^{mn}$ of size mn with entries in $[0, 1]$ (where 0 represents pure black and 1 represents pure white).

The original image $x \in \mathcal{H}$ is assumed to be blurred by a linear operator $L : \mathcal{H} \to \mathcal{H}$ (e.g. the camera is out of focus or in movement during the exposure). Furthermore, it is corrupted with a noise ν, so that only the result $b = Lx + \nu$ is known to us. We want to reconstruct the original image x by considering the minimisation problem

$$\min_{x \in \mathcal{H}} \left(\frac{\mu}{2} \|Lx - b\|^2 + J(Dx)\right),$$

where $\|\cdot\|$ denotes the usual Euclidean norm, $\mu > 0$ is a regularisation parameter, $D : \mathbb{R}^{mn} \to \mathbb{R}^{2mn}$ is the discrete gradient operator given by $Dx = (K_1 x, K_2 x)$, where

$$K_1 : \mathcal{H} \to \mathcal{H}, (K_1 x)_{i,j} := \begin{cases} x_{i+1,j} - x_{i,j}, & i = 1, \dots, m-1; j = 1, \dots, n; \\ 0, & i = m; j = 1, \dots, n \end{cases}$$

$$K_2 : \mathcal{H} \to \mathcal{H}, (K_2 x)_{i,j} := \begin{cases} x_{i,j+1} - x_{i,j}, & i = 1, \dots, m; j = 1, \dots, n-1; \\ 0, & i = 1, \dots, m; j = n, \end{cases}$$

and $J : \mathcal{H} \to \mathbb{R}$ is a regularising functional penalising noisy images. We want to compare several choises of the functional J proposed by [13, 17], all of which have in common that they want to induce sparsity of Dx, i.e. having many components equal to zero.

The *smoothly clipped absolute deviation* (SCAD) penalty was introduced by Fan and Li in [12]. It is defined by

$$\mathrm{SCAD}_{\lambda,a}(z) = \sum_{j=1}^{2mn} g_{\lambda,a}(z_j),$$

where $\lambda > 0$, $a > 1$ and

$$g_{\lambda,a}(z_j) = \begin{cases} \lambda\,|z_j| & \text{if } |z_j| \leq \lambda, \\ \frac{-|z_j|^2+2a\lambda|z_j|-\lambda^2}{2(a-1)} & \text{if } \lambda < |z_j| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |z_j| > a\lambda. \end{cases}$$

$$= \lambda\,|z_j| - \begin{cases} 0 & \text{if } |z_j| \leq \lambda, \\ \frac{(|z_j|-\lambda)^2}{2(a-1)} & \text{if } \lambda < |z_j| \leq a\lambda, \\ \lambda\,|z_j| - \frac{(a+1)\lambda^2}{2} & \text{if } |z_j| > a\lambda. \end{cases}$$

Denoting the part after the curly brace as $h_{\lambda,a}(z_j)$ and $h_{\lambda,a}(z) := \sum_{j=1}^{2mn} h_{\lambda,a}(z_j)$, we have

$$\operatorname{Prox}_{\gamma h^*_{\lambda,a}}(z) = \begin{cases} -\lambda & \text{if } z \leq -(1+\gamma a)\lambda, \\ -\frac{z+\gamma\lambda}{1+\gamma a-\gamma} & \text{if } -(1+\gamma a)\lambda \leq z \leq -\gamma\lambda, \\ 0 & \text{if } |z| \leq \gamma\lambda, \\ \frac{z-\gamma\lambda}{1+\gamma a-\gamma} & \text{if } \gamma\lambda \leq z \leq (1+\gamma a)\lambda, \\ \lambda & \text{if } z \geq (1+\gamma a)\lambda. \end{cases}$$

The *Zhang penalty* [27] is defined by

$$\operatorname{Zhang}_a(z) = \sum_{j=1}^{2mn} g_a(z_j),$$

where $a > 0$ and

$$g_a(z_j) = \begin{cases} \frac{1}{a}\,|z_j| & \text{if } |z_j| < a, \\ 1 & \text{if } |z_j| \geq a. \end{cases}$$

$$= \frac{1}{a}\,|z_j| - \begin{cases} 0 & \text{if } |z_j| < a, \\ \frac{1}{a}(|z_j| - a) & \text{if } |z_j| \geq a. \end{cases}$$

Denoting the part after the curly brace as $h_a(z_j)$ and $h_a(z) := \sum_{j=1}^{2mn} h_a(z_j)$, we have

$$\operatorname{Prox}_{\gamma h^*_a}(z) = \begin{cases} -\frac{1}{a} & \text{if } z \leq -\frac{1}{a} - \gamma a, \\ z + \gamma a & \text{if } -\frac{1}{a} - \gamma a \leq z \leq -\gamma a, \\ 0 & \text{if } -\gamma a \leq z \leq \gamma a, \\ z - \gamma a & \text{if } \gamma a \leq z \leq \frac{1}{a} + \gamma a, \\ \frac{1}{a} & \text{if } z \geq \frac{1}{a} + \gamma a. \end{cases}$$

The *LZOX penalty* [17] is defined by

$$\operatorname{LZOX}_a(z) = \|Dx\|_{\ell^1} - a\,\|Dx\|_{\times},$$

where $\|\cdot\|_{\ell^1}$ denotes (as usual) the sum of the absolute values and

$$\|(u,v)\|_\times := \sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{u_{i,j}^2+v_{i,j}^2},$$

where $y=(u,v)$ is the splitting according to the definition of D. The algorithm (13)–(14) can now be applied to any of the models described above, since the models are written as d. c. problems and the components are easily accessible for computation, with the exception of the function $\|\cdot\|_{\ell^1}\circ D$, see [9]. For the latter, see the following section.

5.1 The proximal point of the anisotropic total variation

In order to apply Algorithm (13)–(14) to any of the problems, we have to calculate the proximal point of the anisotropic total variation by solving the optimization problem

$$\inf\left\{\frac{1}{2\gamma}\|x-b\|^2+\|Dx\|_{\ell^1}\,\middle|\,x\in\mathcal{H}\right\} \tag{26}$$

for some $\gamma>0$ and $b\in\mathcal{H}$ in each step. The Fenchel dual problem [5, Chapter 19] is given by

$$\inf\left\{\frac{\gamma}{2}\|D^*x^*\|^2-\langle b,D^*x^*\rangle\,\middle|\,x^*\in\mathcal{G},\|x^*\|_{\ell^\infty}\le 1\right\}. \tag{27}$$

Instead of solving (26), we could also solve (27) (see [7]), as the following result shows.

Lemma 5. *Let $x^*\in\mathcal{G}$ be a solution of* (27)*. Then $x=b-\gamma D^*x^*$ is a solution of* (26)*.*

Proof. See [5, Example 19.7]. In short:

$$\begin{aligned}0\in D(\gamma D^*x^*-b)+\partial\|\cdot\|_{\ell^1}^*(x^*) &\implies D^*x^*\in D^*\partial\|\cdot\|_{\ell^1}(D(b-\gamma D^*x^*))\\ &\implies \frac{1}{\gamma}(b-x)\in D^*\partial\|\cdot\|_{\ell^1}(Dx)\\ &\iff 0\in\partial\left(\frac{1}{2\gamma}\|(\cdot)-b\|^2+\|D(\cdot)\|_{\ell^1}\right)(x).\end{aligned}$$ □

To the formulation (27), the Forward-Backward algorithm can be applied, since the objective function is differentiable and the feasible set is easy to project on.

5.2 Numerical results

We implemented the FBDC algorithm applied to the model described above and tested the MATLAB code on a PC with Intel Core i5 4670S (4× 3.10GHz) and 8GB DDR3 RAM (1600MHz). Our implementation used the method described in Section 5.1 until the ℓ^∞ distance between two iterations was smaller than 10^{-4}. Both stepsizes were chosen as $\mu_n=\gamma_n=\frac{1}{8\mu}$ for all $n\ge 0$. As initial value, we chose $x_0=b$ and picked $v_0\in\partial h(Kx_0)$.

	$\alpha = 0.00$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
$\mu =$ 1.0	-3.0288	-4.2266	-3.7637	-3.6569	-3.5150	-4.3590	-13.701
$\mu =$ 10.0	5.9227	6.26615	6.414791	6.44871	6.45780	6.28863	4.301090
$\mu =$ 20.0	6.76613	6.90005	**6.93064**	6.917926	6.88018	6.61521	5.305623
$\mu =$ 50.0	6.81752	6.78308	6.65411	6.4923	6.36250	5.780558	4.741993
$\mu =$ 100.0	5.29597	5.23264	5.05189	4.91247	4.739717	4.287092	3.696120
$\mu =$ 200.0	3.088196	3.060511	2.985871	2.930448	2.863122	2.693096	2.477708
$\mu =$ 500.0	1.317390	1.312168	1.298834	1.288983	1.277010	1.246724	1.208036
$\mu =$ 1000.0	0.692487	0.691049	0.687585	0.685057	0.682000	0.674272	0.664401

Table 1: LZOX after 50 iterations

	$\alpha = 0.01$	$\alpha = 0.03$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 1.0$	$\alpha = 3.0$
$\mu =$ 1.0	-43.708	-33.711	-23.148	-13.846	-3.0288	2.4922
$\mu =$ 10.0	-18.781	-9.9406	-3.2070	2.5442	5.9227	**6.97777**
$\mu =$ 20.0	-11.270	-4.8428	0.43533	4.7768	6.76613	6.57299
$\mu =$ 50.0	-4.8333	-1.05553	2.63959	6.46109	6.81752	3.952101
$\mu =$ 100.0	-1.7546	-0.14560	3.16532	6.90202	5.29597	2.129705
$\mu =$ 200.0	-0.41418	0.0619477	2.98543	6.38513	3.088196	1.110186
$\mu =$ 500.0	0.0077144	0.121807	2.101321	3.816813	1.317390	0.482406
$\mu =$ 1000.0	0.0528014	0.127592	1.423684	2.070959	0.692487	0.271777

Table 2: Zhang after 50 iterations

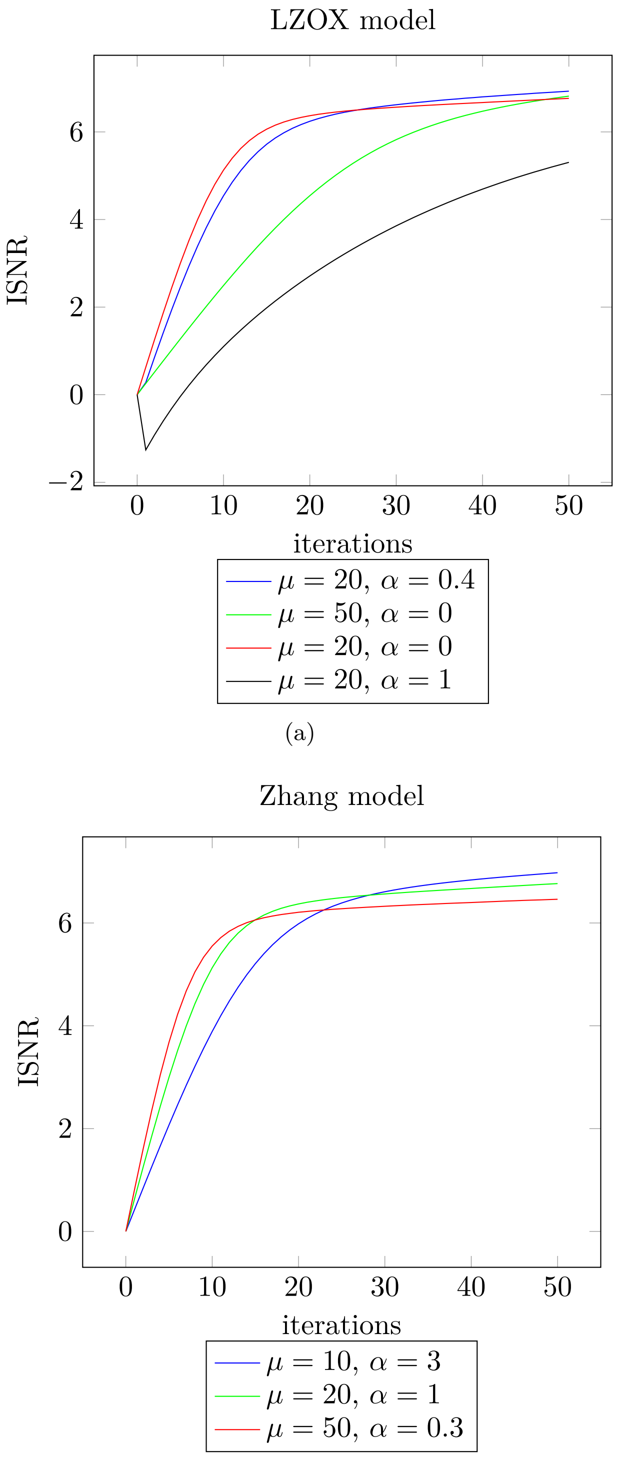
We picked the image `texmos3` from `http://sipi.usc.edu/database/database.php?volume=textures&image=64` and convolved it with a Gaussian kernel with 9 pixels standard devitation. Afterwards we added white noise with standard deviation 50/255, projected the pixels back to the range $[0, 1]$ and saved the image in TIFF format, rounding the brightness values to multiples of 1/255. See Figure 2 for original, blurry and reconstructed image.

The *improvement in signal-to-noise ratio* or *ISNR value* of a reconstruction is given by

$$\mathrm{ISNR}(x_k) = 10 \log_{10} \left(\frac{\|x - b\|^2}{\|x - x_k\|^2} \right),$$

where x is the (usually unknown) original, b is the known blurry and noisy and x_k is the reconstructed image. For the ISNR values after 50 iterations, see Tables 1 and 2. The development of the ISNR values over the iterations is shown in Figure 1.

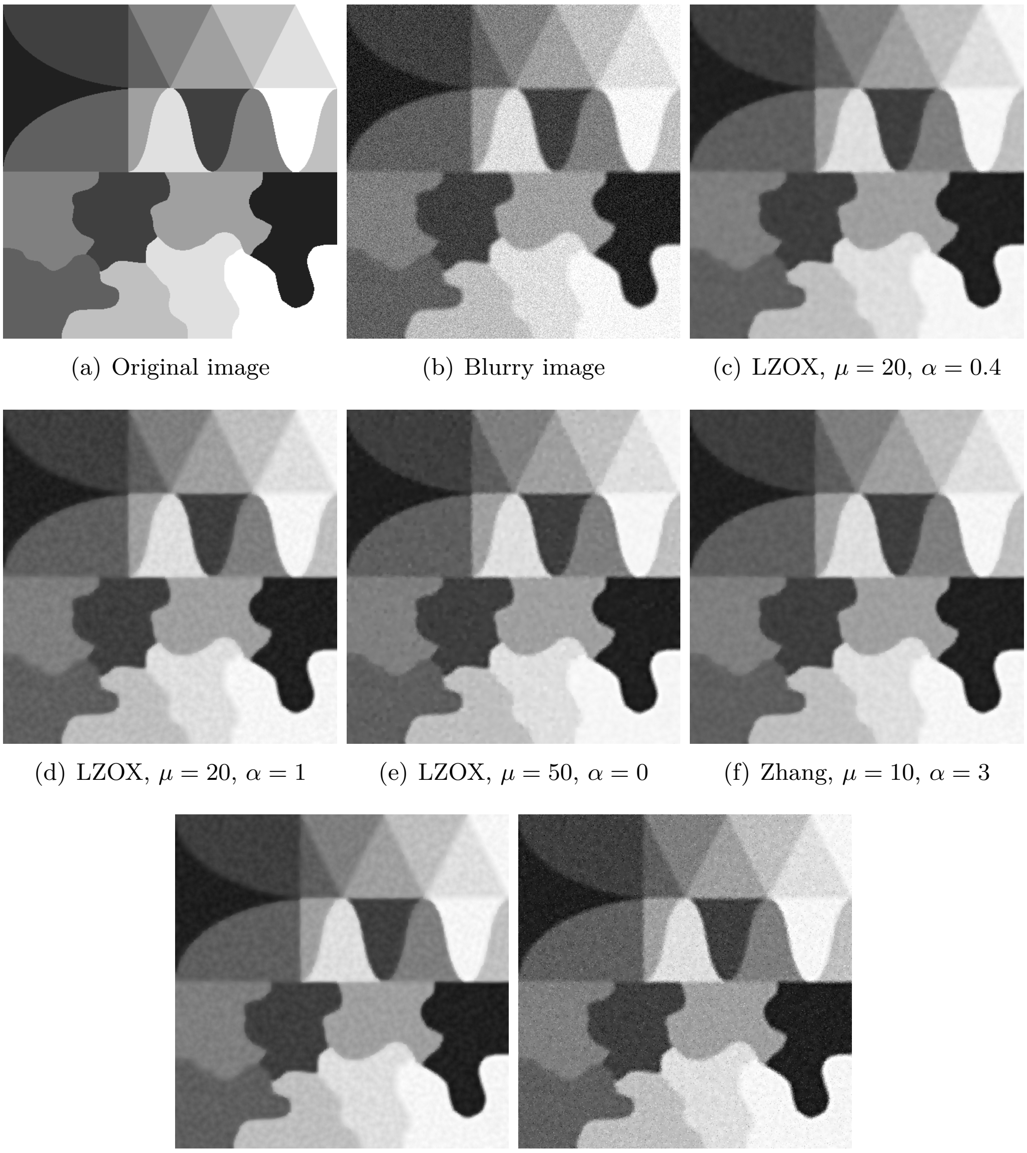
We see that the nonconvex models provide reasonable reconstructions of the original image and the best numerical performance for this particular choice of the stepsizes and the number of iterations is not achieved for the convex model (LZOX with $\alpha = 0$), but for the nonconvex models.



(a)

(b)

Figure 1: ISNR values vs. iterations



(a) Original image

(b) Blurry image

(c) LZOX, $\mu = 20$, $\alpha = 0.4$

(d) LZOX, $\mu = 20$, $\alpha = 1$

(e) LZOX, $\mu = 50$, $\alpha = 0$

(f) Zhang, $\mu = 10$, $\alpha = 3$

(g) Zhang, $\mu = 20$, $\alpha = 1$

(h) Zhang, $\mu = 100$, $\alpha = 0.1$

Figure 2: Original image, blurry and noisy image and reconstructions.

6 Acknowledgements

The authors are grateful to Joseph Salmon for pointing their attraction on the paper [13].

References

[1] Nguyen Thai An, and Nguyen Mau Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *arXiv:1504.08079v4 [math.OC]*, June 2015.

[2] Francisco J. Aragón Artacho, Ronan M.T. Fleming, and Phan T. Vuong. Accelerating the DC algorithm for smooth functions. *arXiv:1507.07375v2 [math.OC]*, September 2016.

[3] Hédy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1–2):5–16, January 2009.

[4] Hédy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129, February 2013.

[5] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* CMS Books in Mathematics. Springer, 2011.

[6] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1–6, January 2014.

[8] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, August 2014.

[9] Radu Ioan Boţ, and Christopher Hendrich. Convergence analysis for a primal-dual monotone + skew splitting algorithm with applications to total variation minimization. *Journal of Mathematical Imaging and Vision*, 49(3):551–568, 2014

[10] Guillaume Carlier. Remarks on Toland's duality, convexity constraint and optimal transport. *Pacific Journal of Optimization*, 4(3):423–432, September 2008.

[11] Bui Van Dinh, Do Sang Kim, and Liguo Jiao. Convergence analysis of algorithms for DC programming. *arXiv:1508.03899v1 [math.OC]*, August 2015.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[13] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, December 2009.

[14] Reiner Horst and Nguyen Van Thoai. DC programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, October 1999.

[15] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier (Grenoble)*, 48(3):769–783, 1998.

[16] Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles.* Éditions du Centre National de la Recherche Scientifique Paris, 87–89, 1963.

[17] Yifei Lou, Tieyong Zeng, Stanley Osher, and Jack Xin. A weighted difference of anisotropic and isotropic total variation model for image processing. *SIAM Journal on Imaging Sciences*, 8(3):1798–1823, 2015.

[18] Paul-Emile Maingé and Abdellatif Moudafi. Convergence of new inertial proximal methods for DC programming. *SIAM Journal on Optimization*, 19(1):397–413, 2008.

[19] Yurii Nesterov. *Introductory Lectures on convex Optimization: A Basic Course*, volume 87 of *Applied Optimization.* Kluwer Academic Publishers, Boston/Dordrecht/London, 2004.

[20] R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series.* Princeton University Press, 1970.

[21] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften.* Springer, 1998.

[22] Wen-yu Sun, Raimundo J.B. Sampaio, and M.A.B. Candido. Proximal point algorithm for minimization of DC function. *Journal of Computational Mathematics*, 21(4):451–462, 2003.

[23] Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to d.c. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

[24] Hoai An Le Thi and Manh Cuong Nguyen. DCA based algorithms for feature selection in multi-class support vector machine. *Annals of Operations Research*, DOI: 10.1007/s10479-016-2333-y, 2016.

[25] John F. Toland. Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66(2):399–415, November 1978.

[26] John F. Toland. A duality principle for non-convex optimisation and the calculus of variations. *Archive for Rational Mechanics and Analysis*, 71(1):41–61, May 1979.

[27] Tong Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, October 2009.

[28] Constantin Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.