Numerical Algorithms manuscript No.
(will be inserted by the editor)

On Starting and Stopping Criteria for Nested Primal-Dual Iterations

Jixin Chen · Ignace Loris

Received: date / Accepted: date

Abstract The importance of an adequate inner loop starting point (as opposed to a sufficient inner loop stopping rule) is discussed in the context of a numerical optimization algorithm consisting of nested primal-dual proximal-gradient iterations. While the number of inner iterations is fixed in advance, convergence of the whole algorithm is still guaranteed by virtue of a warm-start strategy for the inner loop, showing that inner loop "starting rules" can be just as effective as "stopping rules" for guaranteeing convergence. The algorithm itself is applicable to the numerical solution of convex optimization problems

Jixin Chen

East China Normal University (and Université libre de Bruxelles)

Shanghai, China

Jixin.Chen@ulb.ac.be

Ignace Loris, Corresponding author

Université libre de Bruxelles

Brussels, Belgium

igloris@ulb.ac.be

defined by the sum of a differentiable term and two possibly non-differentiable terms. One of the latter terms should take the form of the composition of a linear map and a proximable function, while the differentiable term needs an accessible gradient. The algorithm reduces to the classical proximal gradient algorithm in certain special cases and it also generalizes other existing algorithms. In addition, under some conditions of strong convexity, we show a linear rate of convergence.

Keywords optimization · convergence · forward-backward splitting · primal-dual algorithm · starting rule · stopping rule

Mathematics Subject Classification (2000) 65K10 · 90C06 · 90C25 · 90C90

1 Introduction

Iterative optimization algorithms are based on the availability of simple building blocks related to the cost function that needs minimization. These building blocks, such as e.g. gradients and hessians in case of smooth optimization, should be easy to compute as they are evaluated at every step of the iteration process.

In the framework of the numerical solution of constrained optimization problems of type

$$\min_{u \in C} f(u)$$

with f a real-valued, convex, differentiable function and C a non-empty closed convex set in $\mathbb{R}^d$, projections onto convex sets play a crucial role, such as in the projected gradient algorithm [1]:

$$u_{n+1} = P_C(u_n - \alpha \nabla f(u_n)) \tag{1}$$

(with u_0 arbitrary and $\alpha > 0$ a step length parameter). More generally, the optimization problem

$$\min_{u \in \mathbb{R}^d} f(u) + h(u) \tag{2}$$

(with h a convex proper lower semi-continuous function) can be tackled with the proximal-gradient algorithm (see e.g. [2] and references therein)

$$u_{n+1} = \mathrm{prox}_{\alpha h}(u_n - \alpha \nabla f(u_n)). \tag{3}$$

Apart from the gradient of the differentiable part, here one also needs the *proximal operator* of the non-differentiable function h, which was introduced in [3] (see also Definition 2.2) and for which explicit (and easy to evaluate) expressions exist for several useful cases [2]. One convergence result among many states that the algorithm (3) will converge to a minimizer (if one exists) when $0 < \alpha < 2/L$ where L is the Lipschitz constant of the gradient of f [2].

A large number of generalizations of the proximal gradient algorithm exist, such as e.g. versions with variable metrics [4,5] that exchange the Euclidean distance in the definition of the proximal operator for other ones (e.g. scaled Euclidean distances), depending on the iteration step n (see also [6,7]).

In this paper we are interested in an optimization problem defined by a cost function which consists of three parts instead of two:

$$\min_{u\in\mathbb{R}^d} f(u) + g(Au) + h(u), \tag{4}$$

where f and h are as before and where A is a linear map and g is a convex proper lower semi-continuous function. This type of problem arises naturally when modeling inverse problems, e.g. in image deblurring the function f may represent a quadratic data misfit between blurry data and the unknown image u, the function h may be used to impose positivity of image pixels and the term $g(Au)$ can be used to regularize the inversion by restricting e.g. the total variation of the unknown image (see e.g. [8] for applications of continuous optimization algorithms in imaging). Other applications can be found in the so-called fused lasso problem in statistics [9] or in model selection with grouped variables [10].

If the proximal operator of $\alpha h + \alpha g \circ A$ were available, then algorithm (3) could be used for solving problem (4) (replacing $\text{prox}_{\alpha h}$ in (3) by $\text{prox}_{\alpha h+\alpha g\circ A}$) as follows:

$$u_{n+1} = \text{prox}_{\alpha h+\alpha g\circ A}(u_n - \alpha\,\nabla f(u_n)), \tag{5}$$

(with u_0 arbitrary). However we will assume that the proximal operator of $\alpha h + \alpha g \circ A$ is not explicitly available, rendering algorithm (5) ineffective. Still, we will suppose that the proximal operators $\text{prox}_{\alpha h}$, $\text{prox}_{\alpha g}$ and the linear operator A separately are at our disposal (in many cases of practical interest the latter proximal operators are easier to compute in closed form than

the former one). Such a problem has been studied in [11,12,6,13,14]. Pertinent algorithms have also been deduced from more general splitting algorithms for monotone inclusions [15,16].

It is well-known that the proximal operator appearing in (5) can itself be found using an iterative algorithm based on dual variables (see for e.g. [17–19] for a special case in the area of mathematical imaging). Hence, a nested algorithm (i.e. combining an inner loop with an outer loop) can be a straight-forward way of tackling the described problem. Such nested primal-dual algorithms have already been used in practice for solving large scale optimization problem in mathematical imaging and signal processing [20].

Using an inner loop for the calculation of the proximal operator invariably introduces numerical error in the outer loop. In general, convergence of the proximal gradient algorithm (5) is robust with respect to errors of the proximal operator, in as much as the sum of all errors is finite [2]. However, such a condition is hard to verify in practice. Other (verifiable) conditions have also been proposed [21,22]. The effects of inexact computation on accelerated proximal algorithms have been studied in [23,24].

In this paper we fix the number of inner iterations in advance (thereby completely avoiding the need to check a sufficient inner loop termination condition, while potentially losing control on the accuracy of the approximation of the proximal operator), but use a feedback procedure to guarantee the overall convergence. Therefore the main goal of this paper is to provide a rigorous convergence analysis of a nested iterative algorithm under an a priori finite

termination condition of the inner loop (i.e. number of inner iterations is fixed in advance) with inner loop starting point feedback.

A "warm start" strategy is any method which uses the numerical result of one optimization problem as a starting point for a different, but closely related or perturbed, one. Such strategies are often used (e.g. for computing the solutions of a whole parameter family of optimization problems such as in [25]), but theoretical guarantees or results are lacking. E.g. a warm start strategy is used in an inner loop in the proximal gradient ordered subsets framework applied to computer tomography in [26], but this is only briefly mentioned in an accompanying technical paper [27, Algorithm 4 and below]. Thus nested algorithms have already been proposed in the context of proximal algorithms; however, the convergence analysis presented here is novel, and puts the use of such algorithms on a firmer footing, bypassing a need to rely e.g. on the summability of errors in intermediate computations.

In the following section we will write a specific nested primal-dual algorithm applicable to the described problem. It uses only gradients of the differentiable term, the linear map A (and its transpose) and the proximal operators of g and h. In addition to its convergence we also prove a geometric convergence rate (under the additional assumption of strong convexity). The nested primal-dual algorithm could be interpreted as a generalization of the algorithm proposed in [28] and further developed in [29,30,13] in the sense that it could be identified as corresponding to just a single inner iteration in the algorithm discussed below. In that case, there is no real inner "loop" and the issue of its starting

and stopping rule is absent. In this weak sense, the present proof of convergence generalizes the ones found in [28,13].

2 Nested Primal-Dual Proximal Gradient Algorithm

In the remainder of the paper we assume that $f : \mathbb{R}^d \to \mathbb{R}$ is convex, differentiable and that the gradient of f is Lipschitz continuous (constant L). We also assume that $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^{d'} \to \mathbb{R} \cup \{+\infty\}$ are proper, convex, lower semi-continuous functions. Finally $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a linear map, and $\|A\|$ signifies the largest singular value of the matrix A.

We start by recalling a number of well-known definitions and properties which are necessary for the derivation of the algorithms and for proving their convergence.

Definition 2.1 Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function. The subdifferential of h at the point u is defined as the set

$$\partial h(u) = \{w \in \mathbb{R}^d \quad | \quad h(v) \geq h(u) + \langle w, v - u\rangle \quad \forall v \in \mathbb{R}^d\}. \tag{6}$$

It is easy to see that $\hat{u}$ is a minimizer of h if and only if $0 \in \partial h(\hat{u})$. Also, under mild conditions [31], one can show that $\partial(h_1 + h_2)(u) = \partial h_1(u) + \partial h_2(u)$ and $\partial(g \circ A)(u) = A^T \partial g(Au)$ where A is a linear map.

Definition 2.2 Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function. The proximal operator of h is defined as:

$$\text{prox}_h(a) = \arg\min_{u \in \mathbb{R}^d} \frac{1}{2}\|u - a\|_2^2 + h(u). \tag{7}$$

The proximal operator is a nonexpansive map (Lipschitz continuous with constant 1) defined on all of $\mathbb{R}^d$ [2, Lemma 2.4].

Definition 2.3 Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semicontinuous function. The Fenchel dual of h is is defined as $h^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$:

$$h^*(w) = \sup_u \langle w, u\rangle - h(u). \tag{8}$$

It is again a convex proper lower-semicontinuous function. In fact, on this class, the Fenchel transform is its own inverse: $(h^*)^* = h$ [32, Corollary 12.2.1.].

Next we present some of the classical results of [3] under the form of the following lemmas.

Lemma 2.1 *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semicontinuous function. The following are equivalent:*

1. $u = \operatorname{prox}_{\alpha h}(u + \alpha w) \qquad \forall \alpha > 0$
2. $w \in \partial h(u)$
3. $h(u) + h^*(w) = \langle w, u\rangle$
4. $u \in \partial h^*(w)$
5. $w = \operatorname{prox}_{\beta h^*}(\beta u + w) \qquad \forall \beta > 0$

Furthermore, proximal operators of primal and dual functions h and h^* are related by Moreau's decomposition:

$$\operatorname{prox}_{\alpha h}(u) + \alpha \operatorname{prox}_{\alpha^{-1} h^*}(\alpha^{-1} u) = u \qquad \forall u \in \mathbb{R}^d$$

and any $\alpha > 0$. Therefore it suffices to know $\operatorname{prox}_h(a)$ in order to compute $\operatorname{prox}_{h^*}(a)$ and vice-versa.

Proof See [3]. □

Finally, we will need some further results on the Moreau envelope of a function.

Definition 2.4 Let $\alpha > 0$ and $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex proper and lower semi-continuous. The Moreau envelope of h (of index α) is defined as $\hat{h}_\alpha : \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned}\hat{h}_\alpha(u) &= \min_{v\in\mathbb{R}^d} \tfrac{1}{2\alpha}\|v-u\|_2^2 + h(v) \\ &= \tfrac{1}{2\alpha}\|\operatorname{prox}_{\alpha h}(u) - u\|_2^2 + h(\operatorname{prox}_{\alpha h}(u)).\end{aligned}$$

Lemma 2.2 *The Moreau envelope admits the following properties:*

1. $\hat{h}_\alpha$ *is convex, proper and lower semi-continuous;*
2. $\hat{h}_\alpha$ *is differentiable and* $\nabla \hat{h}_\alpha(u) = \alpha^{-1}(u - \operatorname{prox}_{\alpha h}(u))$ *(Lipschitz, constant* α^{-1}*);*
3. $\hat{h}_\alpha(u) + \widehat{(h^*)}_{1/\alpha}(u/\alpha) = \frac{1}{2\alpha}\|u\|_2^2$ *(hence* $\frac{1}{2\alpha}\|u\|_2^2 - \hat{h}_\alpha(u)$ *is a convex function).*

Proof See [3]. □

Our goal for the remainder of this section is to write an approximate version of the proximal gradient algorithm (5), for problem (4). Informally, it takes the form:

$$u_{n+1} \approx \operatorname{prox}_{\alpha h + \alpha g\circ A}(u_n - \alpha \nabla f(u_n)). \tag{9}$$

In particular, we aim to approximate the proximal operator $\operatorname{prox}_{\alpha h+\alpha g\circ A}$ using an iterative calculation using dual variables. This calculation will involve the proximal operators of αh and of $\beta\alpha^{-1}g^*$, and the linear map A.

Lemma 2.3 *The proximal operator of $\alpha h + \alpha g \circ A$ evaluated at some point a, defined as:*

$$\hat{a} = \text{prox}_{\alpha h + \alpha g \circ A}(a) = \arg\min_{u \in \mathbb{R}^d} \frac{1}{2}\|u - a\|_2^2 + \alpha h(u) + \alpha g(Au),$$

can be computed as $\hat{a} = \text{prox}_{\alpha h}(a - \alpha A^T \hat{v})$ where $\hat{v}$ is the limit of the sequence $(v^k)_{k \in \mathbb{N}}$ defined by the iteration

$$v^{k+1} = \text{prox}_{\beta\alpha^{-1} g^*}(v^k + \beta\alpha^{-1} A \,\text{prox}_{\alpha h}(a - \alpha A^T v^k)), \tag{10}$$

for step size $0 < \beta < 2/\|A\|^2$ and arbitrary v^0.

Proof Writing out the variational equations that determine the minimizer $\hat{a}$, one finds:

$$\hat{a} - a + \alpha\hat{w} + \alpha A^T \hat{v} = 0 \qquad \text{with} \quad \hat{w} \in \partial h(\hat{a}) \quad \text{and} \quad \hat{v} \in \partial g(A\hat{a}).$$

Using Lemma 2.1 the latter inclusions can equivalently be written as (non-linear) equations:

$$\hat{a} = \text{prox}_{\alpha h}(\hat{a} + \alpha\hat{w}) \qquad \hat{v} = \text{prox}_{\beta\alpha^{-1} g^*}(\hat{v} + \beta\alpha^{-1} A\hat{a})$$

where $\beta > 0$ is an arbitrary parameter and g^* is the Fenchel dual of g. In other words $\hat{a}$ and $\hat{v}$ are determined by the equations:

$$\hat{a} = \text{prox}_{\alpha h}(a - \alpha A^T \hat{v}) \qquad \text{and} \qquad \hat{v} = \text{prox}_{\beta\alpha^{-1} g^*}(\hat{v} + \beta\alpha^{-1} A\hat{a})$$

for some $\beta > 0$. Now the variable $\hat{a}$ can be eliminated from the second equation, yielding:

$$\hat{v} = \text{prox}_{\beta\alpha^{-1} g^*}(\hat{v} + \beta\alpha^{-1} A \,\text{prox}_{\alpha h}(a - \alpha A^T \hat{v})).$$

This equation for $\hat{v}$ can be solved using the fixed point iteration (10). Indeed, if we set

$$\varphi(u) = \frac{1}{2\alpha}\|u\|_2^2 - \hat{h}_\alpha(u)$$

(a convex differentiable function according to Lemma 2.2, point 3) and

$$\psi(v) = \alpha^{-1}\varphi(a - \alpha A^T v)$$

(also a convex differentiable function), we see that the gradient of ψ is (Lemma 2.2, point 2):

$$\nabla\,\psi(v) = \alpha^{-1}(-\alpha A)\,\nabla\,\varphi(a - \alpha A^T v) = -\alpha^{-1} A\,\mathrm{prox}_{\alpha h}(a - \alpha A^T v).$$

A Lipschitz constant of the gradient of ψ is $\|A\|^2$. Hence iteration (10) is just the proximal gradient algorithm (3) applied to the "dual problem":

$$\min_v \psi(v) + \alpha^{-1} g^*(v) \tag{11}$$

and therefore converges for $0 < \beta < 2/\|A\|^2$. □

Introducing further auxiliary variables u^k it is also possible to write iteration (10) as:

$$\begin{aligned} &\text{for } k : 0, 1 \ldots \\ &\begin{cases} u^k = \mathrm{prox}_{\alpha h}(a - \alpha A^T v^k) \\ v^{k+1} = \mathrm{prox}_{\beta\alpha^{-1} g^*}(v^k + \beta\alpha^{-1} A u^k) \end{cases} \end{aligned} \tag{12}$$

for step size $0 < \beta < 2/\|A\|^2$ and arbitrary v^0. As the sequence $(u^k)_{k\in\mathbb{N}}$ in (12) converges to $\mathrm{prox}_{\alpha h+\alpha g\circ A}(a)$ (by continuity of $\mathrm{prox}_{\alpha h}$), so does the sequence of averages. One can therefore write the following algorithm for approximating

$\text{prox}_{\alpha h+\alpha g\circ A}(a)$:

$$\begin{aligned}
&\text{for } k : 0 \dots k_{\max} - 1 \\
&\begin{cases} u^k = \text{prox}_{\alpha h}(a - \alpha A^T v^k) \\ v^{k+1} = \text{prox}_{\beta\alpha^{-1} g^*}(v^k + \beta\alpha^{-1} A u^k) \end{cases} \\
&u^{k_{\max}} = \text{prox}_{\alpha h}(a - \alpha A^T v^{k_{\max}}) \\
&\text{prox}_{\alpha h+\alpha g\circ A}(a) \approx \textstyle\sum_{k=1}^{k_{\max}} u^k / k_{\max}
\end{aligned} \tag{13}$$

for some choice of v^0 and $k_{\max} \in \mathbb{N}$.

Instead of imposing an implicit stopping rule on the iteration (13), such as e.g. requiring that $\|v^{k+1} - v^k\|_2 < \epsilon$, we opt to fix the number of iterations $k_{\max}$ in advance. In general, this means that there is no guarantee as to the quality of the approximation (13). Indeed, the starting point could be chosen unfavorably.

If A, A^T, $\text{prox}_{\alpha h}$ and $\text{prox}_{\beta\alpha^{-1} g^*}$ are available, algorithm (13) can be used to compute (an approximation of) the proximal operator present in algorithm (9). By replacing a in (13) by $u_n - \alpha \nabla f(u_n)$ we arrive at Algorithm 1. We will systematically use subscripted n as outer iteration index, and superscripted k as inner iteration index.

It is important to note that, in the proposed nested algorithm, the inner loop starts with the outcome of the previous inner loop: $v_n^0 = v_{n-1}^{k_{\max}}$, and that the number of inner iterations $k_{\max}$ is fixed in advance. It is the former choice, rather than a "sufficient" number of inner iterations, that will allow us to prove convergence of this nested algorithm.

Algorithm 1 Nested primal dual algorithm

Choose u_0, v_0^0, $0 < \alpha < 2/L$, $0 < \beta < 1/\|A\|^2$, $k_{\max} \in \mathbb{N}_0$.

$$
\begin{aligned}
&\text{for } n : 0, 1, \ldots : \\
&\begin{cases}
v_n^0 = v_{n-1}^{k_{\max}} \qquad\qquad \text{for } n > 0 \\
\text{for } k : 0 \ldots k_{\max} - 1 : \\
\quad \begin{cases} u_n^k = \operatorname{prox}_{\alpha h}(u_n - \alpha \nabla f(u_n) - \alpha A^T v_n^k) \\ v_n^{k+1} = \operatorname{prox}_{\beta\alpha^{-1} g^*}\left(v_n^k + \beta\alpha^{-1} A u_n^k\right) \end{cases} \\
u_n^{k_{\max}} = \operatorname{prox}_{\alpha h}(u_n - \alpha \nabla f(u_n) - \alpha A^T v_n^{k_{\max}}) \\
u_{n+1} = \sum_{k=1}^{k_{\max}} u_n^k / k_{\max}
\end{cases}
\end{aligned}
\tag{14}
$$

We remark that all iterates u_n ($n \geq 1$) in Algorithm 1 are in the domain of h, but not necessarily in the domain of $g \circ A$. In the special case $h = 0$ and $k_{\max} = 1$ (just one inner iteration) Algorithm 1 reduces to

$$
\begin{cases}
z_n = u_n - \alpha \nabla f(u_n) \\
v_{n+1} = \operatorname{prox}_{\beta\alpha^{-1} g^*}\left(v_n + \beta\alpha^{-1} A(z_n - \alpha A^T v_n)\right) \\
u_{n+1} = z_n - \alpha A^T v_{n+1},
\end{cases}
\tag{15}
$$

which was proposed in [28] and further studied in [29,13]. It was also interpreted in [7] (see also [6,33]) as a special case of a novel scheme extending several classical ones, like the forward–backward and Douglas–Rachford methods, as well as the more recent algorithm of Chambolle and Pock [34]. Later, it was re-derived under the name "Proximal Alternating Predictor–Corrector" (PAPC) algorithm in [30].

If a minimizer to problem (4) exists, algorithm (15) converges for $0 < \alpha < 2/L$ and $0 < \beta < 1/\|A\|^2$ [28,29]. In the following section, we will prove convergence of Algorithm 1 under the same conditions.

Algorithm 1 is very similar to the one used in [20] for the special case of so-called Total Variation image denoising and deblurring problems. The main difference lies in the absence (in [20]) of a feedback strategy for the inner loop: the authors of [20] restart the inner iteration at $v_n^0 = 0$ (for all n) and observe that this, in combination with a fixed number of inner iterations, may lead to non-convergence of the outer loop. The main contribution of this paper therefore is the convergence resulting from the feedback strategy $v_n^0 = v_{n-1}^{k_{\max}}$ (for all $n > 0$).

3 Convergence Results

Three further lemmas are needed for proving convergence of Algorithm 1.

Lemma 3.1 *The minimizers $\hat{u}$ of problem (4) are characterized by the equations*

$$\begin{cases} \hat{u} = \operatorname{prox}_{\alpha h}(\hat{u} - \alpha \nabla f(\hat{u}) - \alpha A^T \hat{v}) \\ \hat{v} = \operatorname{prox}_{\beta\alpha^{-1} g^*}\left(\hat{v} + \beta\alpha^{-1} A\hat{u}\right) \end{cases} \tag{16}$$

for any $\alpha, \beta > 0$.

Proof The minimizers of (4) are characterized by the inclusion $0 \in \partial(f + h + g \circ A)(\hat{u})$, or

$$0 = \nabla f(\hat{u}) + \hat{w} + A^T \hat{v} \qquad \text{with} \quad \hat{w} \in \partial h(\hat{u}) \quad \text{and} \quad \hat{v} \in \partial g(A\hat{u}). \tag{17}$$

The two latter inclusions can also be written as (see Lemma 2.1):

$$\hat{u} = \mathrm{prox}_{\alpha h}(\hat{u} + \alpha\hat{w}) \quad \text{and} \quad \hat{v} = \mathrm{prox}_{\beta\alpha^{-1}g^*}(\hat{v} + \beta\alpha^{-1}A\hat{u})$$

where $\alpha, \beta > 0$ are arbitrary parameter and g^* is the Fenchel dual of g. One obtains equations (16) by using the first equation of (17) to eliminate $\hat{w}$. □

Lemma 3.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function with Lipschitz continuous gradient (constant L). It follows that $L^{-1}\nabla f$ is firmly non-expansive:*

$$\|\nabla f(u) - \nabla f(v)\|_2^2 \leq L\langle \nabla f(u) - \nabla f(v), u - v\rangle \qquad \forall u, v \in \mathbb{R}^d. \tag{18}$$

Proof See [35, Part 2, Chapter X, Th. 4.2.2]. □

Lemma 3.3 *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function. The equality $x^+ = \mathrm{prox}_h(x^- + \Delta)$ is equivalent to the inequality:*

$$\|x^+ - x\|_2^2 \leq \|x^- - x\|_2^2 - \|x^+ - x^-\|_2^2 + 2\langle x^+ - x, \Delta\rangle + 2h(x) - 2h(x^+)$$

for all $x \in \mathbb{R}^d$.

Proof $x^+ = \mathrm{prox}_h(x^- + \Delta)$

$\Leftrightarrow \quad x^+ = \arg\min_x \frac{1}{2}\|x - (x^- + \Delta)\|_2^2 + h(x)$

$\Leftrightarrow \quad 0 \in x^+ - x^- - \Delta + \partial h(x^+)$

$\Leftrightarrow \quad x^- + \Delta - x^+ \in \partial h(x^+)$

$\Leftrightarrow \quad h(x) \geq h(x^+) + \langle x^- + \Delta - x^+, x - x^+\rangle \qquad \forall x$

$\Leftrightarrow \quad \|x^+ - x\|_2^2 \leq \|x^- - x\|_2^2 - \|x^+ - x^-\|_2^2 + 2\langle x^+ - x, \Delta\rangle + 2h(x) - 2h(x^+)$

□

We are now ready to state and prove the main theorems.

Theorem 3.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function with Lipschitz continuous gradient (constant L), $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function, $g : \mathbb{R}^{d'} \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function and let $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a linear map.*

Let $0 < \alpha < 2/L$, $0 < \beta < 1/\|A\|^2$ and $k_{\max} \in \mathbb{N}_0$. If the optimization problem (4) admits a solution, the nested primal-dual Algorithm 1 will converge to one.

Proof Let $\hat{u} \in \arg\min_u f(u) + g(Au) + h(u)$, i.e. there exists $\hat{v}$ such that equations (16) are satisfied.

We use Lemma 3.3 with $x = \hat{u}$ and the definition of $u_n^{k_{\max}}$ in Algorithm 1:

$$\begin{aligned}\|u_n^{k_{\max}} - \hat{u}\|_2^2 \leq \|u_n - \hat{u}\|_2^2 - \|u_n^{k_{\max}} - u_n\|_2^2 + 2\alpha h(\hat{u}) - 2\alpha h(u_n^{k_{\max}}) \\ -2\alpha\langle u_n^{k_{\max}} - \hat{u}, \nabla f(u_n) + A^T v_n^{k_{\max}}\rangle,\end{aligned} \tag{19}$$

on the definition of u_n^k in Algorithm 1 (and choosing $x = u_n^{k+1}$):

$$\begin{aligned}\|u_n^k - u_n^{k+1}\|_2^2 \leq \|u_n - u_n^{k+1}\|_2^2 - \|u_n^k - u_n\|_2^2 + 2\alpha h(u_n^{k+1}) - 2\alpha h(u_n^k) \\ -2\alpha\langle u_n^k - u_n^{k+1}, \nabla f(u_n) + A^T v_n^k\rangle\end{aligned} \tag{20}$$

for $k : 0 \ldots k_{\max} - 1$, and on the first line of equations (16) with $x = u_n^0$:

$$\begin{aligned}\|\hat{u} - u_n^0\|_2^2 \leq \|\hat{u} - u_n^0\|_2^2 - \|\hat{u} - \hat{u}\|_2^2 - 2\alpha\langle \hat{u} - u_n^0, \nabla f(\hat{u}) + A^T\hat{v}\rangle \\ +2\alpha h(u_n^0) - 2\alpha h(\hat{u}).\end{aligned} \tag{21}$$

Applying Lemma 3.3 again to the definition of u_n^k in Algorithm 1 (now with $x = \hat{u}$):

$$\|u_n^k - \hat{u}\|_2^2 \leq \|u_n - \hat{u}\|_2^2 - \|u_n^k - u_n\|_2^2 - 2\alpha\langle u_n^k - \hat{u}, \nabla f(u_n) + A^T v_n^k\rangle$$

$$+2\alpha h(\hat{u}) - 2\alpha h(u_n^k)$$

for $k : 1 \dots k_{\max} - 1$ and to the first line of equations (16) (with $x = u_n^k$)

$$\|\hat{u} - u_n^k\|_2^2 \leq \|\hat{u} - u_n^k\|_2^2 - \|\hat{u} - \hat{u}\|_2^2 - 2\alpha\langle \hat{u} - u_n^k, \nabla f(\hat{u}) + A^T \hat{v}\rangle$$

$$+2\alpha h(u_n^k) - 2\alpha h(\hat{u})$$

together yields:

$$\begin{aligned}\|u_n^k - \hat{u}\|_2^2 \leq \|u_n - \hat{u}\|_2^2 - \|u_n^k - u_n\|_2^2 - 2\alpha\langle u_n^k - \hat{u}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\ -2\alpha\langle u_n^k - \hat{u}, A^T(v_n^k - \hat{v})\rangle\end{aligned} \tag{22}$$

for $k : 1 \dots k_{\max} - 1$.

Finally, we apply Lemma 3.3 to the definition of v_n^{k+1} in Algorithm 1 (with $x = \hat{v}$):

$$\|v_n^{k+1} - \hat{v}\|_2^2 \leq \|v_n^k - \hat{v}\|_2^2 - \|v_n^{k+1} - v_n^k\|_2^2 + 2\beta\alpha^{-1}\langle v_n^{k+1} - \hat{v}, Au_n^k\rangle$$

$$+2\beta\alpha^{-1}g^*(\hat{v}) - 2\beta\alpha^{-1}g^*(v_n^{k+1})$$

and to the second equation in system (16) (with $x = v_n^{k+1}$):

$$\|\hat{v} - v_n^{k+1}\|_2^2 \leq \|\hat{v} - v_n^{k+1}\|_2^2 - \|\hat{v} - \hat{v}\|_2^2 + 2\beta\alpha^{-1}\langle \hat{v} - v_n^{k+1}, A\hat{u}\rangle$$

$$+2\beta\alpha^{-1}g^*(v_n^{k+1}) - 2\beta\alpha^{-1}g^*(\hat{v})$$

which together give:

$$\|v_n^{k+1} - \hat{v}\|_2^2 \leq \|v_n^k - \hat{v}\|_2^2 - \|v_n^{k+1} - v_n^k\|_2^2 + 2\beta\alpha^{-1}\langle v_n^{k+1} - \hat{v}, A(u_n^k - \hat{u})\rangle \tag{23}$$

for $k : 0 \dots k_{\max} - 1$.

By adding β times the inequalities (19), (20) over $k : 0 \dots k_{\max}-1$, (21) and (22) over $k : 1 \dots k_{\max}-1$ and α^2 times inequalities (23) over $k : 0 \dots k_{\max}-1$, and after canceling some (but not yet all) terms, one has:

$$\begin{aligned}
&\textstyle\sum_{k=1}^{k_{\max}} \beta\|u_n^k - \hat{u}\|_2^2 + \sum_{k=0}^{k_{\max}-1} \alpha^2\|v_n^{k+1} - \hat{v}\|_2^2 \leq \sum_{k=0}^{k_{\max}-1} \Big(\beta\|u_n - \hat{u}\|_2^2 \\
&-\beta\|u_n^k - u_n^{k+1}\|_2^2 - \beta\|u_n^k - u_n\|_2^2 - 2\alpha\beta\langle u_n^k - \hat{u}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\
&+\alpha^2\|v_n^k - \hat{v}\|_2^2 - \alpha^2\|v_n^{k+1} - v_n^k\|_2^2\Big) \\
&-2\alpha\beta\langle u_n^{k_{\max}} - \hat{u}, A^T v_n^{k_{\max}}\rangle - 2\alpha\beta \textstyle\sum_{k=0}^{k_{\max}-1}\langle u_n^k - u_n^{k+1}, A^T v_n^k\rangle \\
&-2\alpha\beta\langle \hat{u} - u_n^0, A^T\hat{v}\rangle - 2\alpha\beta \textstyle\sum_{k=1}^{k_{\max}-1}\langle u_n^k - \hat{u}, A^T(v_n^k - \hat{v})\rangle \\
&+2\alpha\beta \textstyle\sum_{k=0}^{k_{\max}-1}\langle v_n^{k+1} - \hat{v}, A(u_n^k - \hat{u})\rangle.
\end{aligned} \tag{24}$$

On the last three lines of inequality (24), all terms proportional to $\hat{u}$ or $\hat{v}$ drop, and the remaining inner products expand to:

$$\begin{aligned}
&-\langle u_n^{k_{\max}}, A^T v_n^{k_{\max}}\rangle - \textstyle\sum_{k=0}^{k_{\max}-1}\langle u_n^k, A^T v_n^k\rangle + \sum_{k=0}^{k_{\max}-1}\langle u_n^{k+1}, A^T v_n^k\rangle \\
&- \textstyle\sum_{k=1}^{k_{\max}-1}\langle u_n^k, A^T v_n^k\rangle + \sum_{k=0}^{k_{\max}-1}\langle v_n^{k+1}, Au_n^k\rangle
\end{aligned}$$

which equals:

$$\sum_{k=0}^{k_{\max}-1} \langle u_n^k - u_n^{k+1}, A^T(v_n^{k+1} - v_n^k)\rangle.$$

Hence one obtains from expression (24):

$$\begin{aligned}
&\textstyle\sum_{k=1}^{k_{\max}} \beta\|u_n^k - \hat{u}\|_2^2 + \sum_{k=0}^{k_{\max}-1} \alpha^2\|v_n^{k+1} - \hat{v}\|_2^2 \leq \sum_{k=0}^{k_{\max}-1} \Big(\beta\|u_n - \hat{u}\|_2^2 \\
&-\beta\|u_n^k - u_n^{k+1}\|_2^2 - \beta\|u_n^k - u_n\|_2^2 - 2\alpha\beta\langle u_n^k - \hat{u}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\
&+\alpha^2\|v_n^k - \hat{v}\|_2^2 - \alpha^2\|v_n^{k+1} - v_n^k\|_2^2 + 2\alpha\beta\langle u_n^k - u_n^{k+1}, A^T(v_n^{k+1} - v_n^k)\rangle\Big).
\end{aligned} \tag{25}$$

On the second line of (25) one can use the following bound:

$$
\begin{aligned}
&-\|u_n^k - u_n\|_2^2 - 2\alpha\langle u_n^k - \hat{u}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\
&= \alpha^2 \|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 - 2\alpha\langle u_n - \hat{u}, \nabla f(u_n) - \nabla f(\hat{u})\rangle \\
&\qquad -\|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2 \\
&\leq \alpha(\alpha - 2/L)\|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 - \|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2
\end{aligned}
$$

where we have used the fact that $L^{-1}\nabla f$ is firmly non expansive (Lemma 3.2), while the scalar products on the last line of inequality (25) can be replaced by

$$
\begin{aligned}
2\alpha\beta\langle u_n^k - u_n^{k+1}, A^T(v_n^{k+1} - v_n^k)\rangle &= \beta\|u_n^k - u_n^{k+1}\|_2^2 + \beta\alpha^2\|A^T(v_n^{k+1} - v_n^k)\|_2^2 \\
&\quad -\beta\|u_n^k - u_n^{k+1} - \alpha A^T(v_n^{k+1} - v_n^k)\|_2^2.
\end{aligned} \tag{26}
$$

Hence we can deduce that

$$
\begin{aligned}
&\textstyle\sum_{k=0}^{k_{\max}-1} \left(\beta\|u_n^{k+1} - \hat{u}\|_2^2 + \alpha^2\|v_n^{k+1} - \hat{v}\|_2^2\right) \\
&\leq \textstyle\sum_{k=0}^{k_{\max}-1} \Big(\beta\|u_n - \hat{u}\|_2^2 + \alpha^2\|v_n^k - \hat{v}\|_2^2 + \alpha\beta(\alpha - 2/L)\|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 \\
&-\beta\|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2 - \alpha^2\|v_n^{k+1} - v_n^k\|_2^2 \\
&+\alpha^2\beta\|A^T(v_n^{k+1} - v_n^k)\|_2^2 - \beta\|u_n^k - u_n^{k+1} - \alpha A^T(v_n^{k+1} - v_n^k)\|_2^2\Big).
\end{aligned} \tag{27}
$$

Now we use the convexity of $\|u_{n+1} - \hat{u}\|_2^2$ (as a function of u_{n+1}) and the last line of Algorithm 1 to write:

$$
\|u_{n+1} - \hat{u}\|_2^2 \leq \gamma \sum_{k=0}^{k_{\max}-1} \|u_n^{k+1} - \hat{u}\|_2^2
$$

where $\gamma = 1/k_{\max}$. Together with inequality (27), and using the relation $v_{n+1}^0 = v_n^{k_{\max}}$, we finally find:

$$\beta\|u_{n+1} - \hat{u}\|_2^2 + \alpha^2\gamma\|v_{n+1}^0 - \hat{v}\|_2^2 \leq \beta\|u_n - \hat{u}\|_2^2 + \alpha^2\gamma\|v_n^0 - \hat{v}\|_2^2$$

$$+\alpha\beta(\alpha - 2/L)\|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 - \beta\gamma\textstyle\sum_{k=0}^{k_{\max}-1}\|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2$$

$$-\alpha^2\gamma\textstyle\sum_{k=0}^{k_{\max}-1}\|v_n^{k+1} - v_n^k\|_A^2 - \beta\gamma\sum_{k=0}^{k_{\max}-1}\|u_n^k - u_n^{k+1} - \alpha A^T(v_n^{k+1} - v_n^k)\|_2^2 \tag{28}$$

where we have used the norm $\|v\|_A^2 = \|v\|_2^2 - \beta\|A^T v\|_2^2$ (it is a norm because $0 < \beta < 1/\|A\|^2$).

Relation (28) and assumption $0 < \alpha < 2/L$ implies that the sequence $(u_n, v_n^0)_{n\in\mathbb{N}}$ is bounded. Hence a limit point exists: $(u_{n_j}, v_{n_j}^0) \overset{j\to\infty}{\longrightarrow} (u^\dagger, v^\dagger)$. By summing inequalities (28) from $n = 0$ until $n = N$ one deduces also that:

$$\|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 \overset{n\to\infty}{\longrightarrow} 0,$$

$$\|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2 \overset{n\to\infty}{\longrightarrow} 0 \qquad k : 0 \ldots k_{\max} - 1,$$

$$\|v_n^{k+1} - v_n^k\|_A^2 \overset{n\to\infty}{\longrightarrow} 0 \qquad k : 0 \ldots k_{\max} - 1,$$

$$\|u_n^k - u_n^{k+1} - \alpha A^T(v_n^{k+1} - v_n^k)\|_2^2 \overset{n\to\infty}{\longrightarrow} 0 \qquad k : 0 \ldots k_{\max} - 1.$$

This in turn implies that

$$u_{n_j}^{k+1} \overset{j\to\infty}{\longrightarrow} u^\dagger, \qquad v_{n_j}^{k+1} \overset{j\to\infty}{\longrightarrow} v^\dagger, \qquad k : 0 \ldots k_{\max} - 1$$

also. It follows from the continuity of the operations in the right hand sides of Algorithm 1 that $(u^\dagger, v^\dagger)$ satisfies the equations (16), which characterize the minimizers of problem (4). One can then replace $(\hat{u}, \hat{v})$ by $(u^\dagger, v^\dagger)$ in inequality

(28) to obtain

$$\beta\|u_{n+1} - u^\dagger\|_2^2 + \alpha^2\gamma\|v_{n+1}^0 - v^\dagger\|_2^2 \leq \beta\|u_n - u^\dagger\|_2^2 + \alpha^2\gamma\|v_n^0 - v^\dagger\|_2^2$$

This then implies the convergence of the whole sequence (u_n, v_n^0) to $(u^\dagger, v^\dagger)$.

□

As usual, one expects that better convergence results can be obtained when strong convexity of the objective function is assumed. In [31, Example 27.12] the linear convergence rate (to the unique minimizer $\hat{u}$ of problem (2)) of the proximal-gradient algorithm (3) is proven, when f is strongly convex (parameter μ), ∇f is Lipschitz continuous (parameter L) and $0 < \alpha < 2/L$:

$$\|u_{n+1} - \hat{u}\|_2^2 \leq (1 + \mu\alpha(\alpha L - 2))\|u_n - \hat{u}\|_2^2,$$

where $0 \leq 1 + \mu\alpha(\alpha L - 2) < 1$. In [36] a linear convergence rate of the proximal gradient algorithm (3) is shown for strongly convex h (instead of f). The following theorem thus complements the results of [28,29,13,36].

Theorem 3.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function with Lipschitz continuous gradient (constant L), $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function, $g : \mathbb{R}^{d'} \to \mathbb{R} \cup \{+\infty\}$ be a convex, proper, lower semi-continuous function and let $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a linear map. Let $0 < \alpha < 2/L$, $0 < \beta < 1/\|A\|^2$ and $k_{\max} \in \mathbb{N}_0$. In addition, we assume that f is strongly convex (parameter μ), that $h = 0$ and that A^T is coercive (parameter $\sigma > 0$): $\|A^T v\|_2 \geq \sigma\|v\|_2$ for all $v \in \mathbb{R}^{d'}$.*

Then the primal-dual Algorithm 1 converges to the minimizer $\hat{u}$ of problem (4) at a linear linear rate:

$$\beta\|u_{n+1}-\hat{u}\|_2^2+\alpha^2\gamma\|v_{n+1}^0-\hat{v}\|_2^2 \leq \epsilon\left(\beta\|u_n-\hat{u}\|_2^2+\alpha^2\gamma\|v_n^0-\hat{v}\|_2^2\right)$$

for some $0 \leq \epsilon < 1$ and $\gamma = 1/k_{\max}$. Here, $\hat{v}$ is the dual variable of equations (16).

Proof As f is strongly convex, problem (4) is guaranteed to have a (unique) solution $\hat{u}$. Hence, there exists $\hat{v}$ such that equations (16) are satisfied.

We start from inequality (25) derived in the proof of Theorem 3.1. We use the following bound:

$$\begin{aligned}
&\|u_n-\hat{u}\|_2^2-\|u_n^k-u_n\|_2^2+2\alpha\langle u_n^k-\hat{u},\nabla f(\hat{u})-\nabla f(u_n)\rangle \\
&=\|u_n-\hat{u}\|_2^2+\alpha^2\|\nabla f(u_n)-\nabla f(\hat{u})\|_2^2-2\alpha\langle u_n-\hat{u},\nabla f(u_n)-\nabla f(\hat{u})\rangle \\
&\qquad -\|u_n^k-u_n+\alpha(\nabla f(u_n)-\nabla f(\hat{u}))\|_2^2 \\
&\leq\|u_n-\hat{u}\|_2^2+(\alpha^2L-2\alpha)\langle u_n-\hat{u},\nabla f(u_n)-\nabla f(\hat{u})\rangle \\
&\qquad -\|u_n^k-u_n+\alpha(\nabla f(u_n)-\nabla f(\hat{u}))\|_2^2 \\
&\leq(1+\mu\alpha(\alpha L-2))\|u_n-\hat{u}\|_2^2-\|u_n^k-u_n+\alpha(\nabla f(u_n)-\nabla f(\hat{u}))\|_2^2
\end{aligned}$$

where we have used the fact that $L^{-1}\nabla f$ is firmly non expansive (lemma 3.2), that ∇f is strongly monotone

$$\langle\nabla f(u_n)-\nabla f(\hat{u}),u_n-\hat{u}\rangle \geq \mu\|u_n-\hat{u}\|_2^2,$$

(a consequence of the strong convexity of f, [37, Theorem 5.24]) and the assumption that $\alpha L-2<0$. The second term on the right hand side can be

written as:

$$\|u_n^k - u_n + \alpha(\nabla f(u_n) - \nabla f(\hat{u}))\|_2^2 = \alpha^2 \|A^T(v_n^k - \hat{v})\|_2^2$$

on account of the assumption that $h = 0$ (in this case $\text{prox}_{\alpha h}$ is the identity), the definition of u_n^k in Algorithm 1 and the first line in equations (16).

We thus find from inequality (25):

$$\sum_{k=0}^{k_{\max}-1} \left(\beta \|u_n^{k+1} - \hat{u}\|_2^2 + \alpha^2 \|v_n^{k+1} - \hat{v}\|_2^2 \right)$$

$$\leq \sum_{k=0}^{k_{\max}-1} \Big(\beta(1 + \mu\alpha(\alpha L - 2)) \|u_n - \hat{u}\|_2^2 - \alpha^2 \beta \|A^T(v_n^k - \hat{v})\|_2^2 - \beta \|u_n^k - u_n^{k+1}\|_2^2$$

$$+\alpha^2 \|v_n^k - \hat{v}\|_2^2 - \alpha^2 \|v_n^{k+1} - v_n^k\|_2^2 + 2\alpha\beta \langle u_n^k - u_n^{k+1}, A^T(v_n^{k+1} - v_n^k) \rangle \Big) .$$

The inner products on the last line can be bounded using the first two terms in the right hand side of expression (26), such that one finds:

$$\sum_{k=0}^{k_{\max}-1} \left(\beta \|u_n^{k+1} - \hat{u}\|_2^2 + \alpha^2 \|v_n^{k+1} - \hat{v}\|_2^2 \right)$$

$$\leq \sum_{k=0}^{k_{\max}-1} \left(\beta(1 + \mu\alpha(\alpha L - 2)) \|u_n - \hat{u}\|_2^2 + \alpha^2 \|v_n^k - \hat{v}\|_A^2 - \alpha^2 \|v_n^{k+1} - v_n^k\|_A^2 \right)$$

using the norm $\|v\|_A^2 = \|v\|_2^2 - \beta \|A^T v\|_2^2$. The last term on the right hand side is dropped and the definition $u_{n+1} = \gamma \sum_{k=0}^{k_{\max}-1} u_n^{k+1}$ (with $\gamma = k_{\max}^{-1}$) and the convexity of $\| \cdot - \hat{u}\|_2^2$ then imply that:

$$\beta \|u_{n+1} - \hat{u}\|_2^2 + \sum_{k=0}^{k_{\max}-1} \alpha^2 \gamma \|v_n^{k+1} - \hat{v}\|_2^2 \leq$$

$$\beta(1 + \mu\alpha(\alpha L - 2)) \|u_n - \hat{u}\|_2^2 + \sum_{k=0}^{k_{\max}-1} \alpha^2 \gamma \|v_n^k - \hat{v}\|_A^2 .$$

As A^T is coercive, one has that $\|v_n^k - \hat{v}\|_A^2 \leq (1 - \beta\sigma^2) \|v_n^k - \hat{v}\|_2^2$ such that:

$$\beta \|u_{n+1} - \hat{u}\|_2^2 + \alpha^2 \gamma \|v_{n+1}^0 - \hat{v}\|_2^2 \leq$$

$$\beta(1 + \mu\alpha(\alpha L - 2)) \|u_n - \hat{u}\|_2^2 + \alpha^2 \gamma (1 - \beta\sigma^2) \|v_n^0 - \hat{v}\|_2^2$$

as $v_{n+1}^0 = v_n^{k_{\max}}$. By setting $\epsilon = \max((1+\mu\alpha(\alpha L-2)), 1-\beta\sigma^2)$ one finds the announced inequality.

In order to show that $0 \leq \epsilon < 1$, one proceeds as follows. On the one hand, $\alpha L - 2 < 0$ implies $1+\mu\alpha(\alpha L-2) < 1$, while $1+\mu\alpha(\alpha L-2)$ reaches a minimum for $\alpha = 1/L$. This minimum is $1 - \mu/L \geq 0$ as $\mu \leq L$ for strongly convex functions f (with parameter μ) with Lipschitz continuous gradient (parameter L). One sees that $0 < 1 - \beta\sigma^2 < 1$ on account of $0 < \sigma^2 \leq \|A\|^2 < \beta^{-1}$. □

The proof of Theorem 3.2 unfortunately requires the assumption that $h = 0$. When $h = 0$ the dual problem (11) reduces to a quadratic plus proximable term:

$$\min_v \frac{1}{2}\|a/\alpha - A^T v\|_2^2 + \frac{1}{\alpha} g^*(v).$$

When A^T is coercive, the first term is strongly convex. For general h we conjecture that a linear convergence rate still holds for Algorithm 1 when one assumes, in addition to the strong convexity of f, that the function ψ appearing in the dual problem (11) is strongly convex. In [13] a linear convergence rate is shown for Algorithm 1 with $k_{\max} = 1$ and assuming that g^* is strongly convex (in addition to some other assumptions on f and A).

4 Conclusions

A generalization of the proximal gradient algorithm (3) consisting of nested primal and dual iterations was discussed and convergence was shown. The iterative algorithm requires access to a gradient and two proximal operators, but not to the inverse of the linear operator appearing in problem (4). Similar

problems and related algorithms are also discussed in [11,7]. Under some additional conditions (related to strong convexity of the cost function) a linear convergence rate was shown.

Nested iterative algorithms are abundant in numerical and applied mathematics. The proposed algorithm is very similar to the one in [20]. The main novelty lies in the rigorous discussion of the inner loop starting and stopping criterion. One often encounters inner loop stopping criteria of the form $\|v_n^k - v_n^{k+1}\|_2 < \epsilon$, which may give satisfactory numerical results, but may not guarantee convergence of the outer loop. In addition, such a condition may just indicate slow convergence of the inner loop. Additionally, the inner loop starting point is often neglected in theoretical descriptions. In practice, a feedback/warm start mechanism of type $v_n^0 = v_{n-1}^{k_{\max}}$ is sometimes added to "speed up" convergence of the inner loop. Here we have shown that such a small change can already be sufficient to guarantee convergence. Such a discussion has not been given before, and it is the main contribution of this paper: in the context of nested iterative algorithms, inner loop "starting rules" can be just as effective as "stopping rules" for guaranteeing convergence.

No numerical experiments are presented. In fact, the proposed Algorithm 1 cannot be expected to be state-of-the-art by itself (lack of variable step length or line-search strategies [5,38]). The point here is just to prove that the described mechanism is sufficient for convergence. More sophisticated algorithms, e.g. incorporating line-search rules to speed-up convergence, exist. In those

cases too, one could investigate the role of the warm start mechanism on the convergence of nested iterations.

Another possible extension concerns the convergence of nested *accelerated* primal-dual algorithms of Nesterov type. The convergence of the iterates of accelerated projected-gradient [39] and proximal-gradient algorithms [40] was shown in [41]. Nested primal-dual versions were proposed in [20], again without feedback. The proof of convergence of algorithms of that type is still an open problem.

Another generalization concerns the use of variable stepsizes (α_n instead of α and β_n instead of β) in Algorithm 1. Finally, the condition $0 < \beta < 1/\|A\|^2$ seems to be too restrictive in view of the step size condition in Lemma 2.3 ($0 < \beta < 2/\|A\|^2$). A variation of Algorithm 1 with a different feedback strategy is described in [42].

Acknowledgements JC is sponsored by the China Scholarship Council. IL is a Research Associate of the Fonds de la Recherche Scientifique - FNRS and is also supported by a ULB ARC grant.

References

1. Goldstein, A.A.: Convex programming in Hilbert space. Bulletin of the American Mathematical Society **70**, 709–710 (1964)
2. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005). DOI 10.1137/050626090
3. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)

4. Combettes, P.L., Vũ, B.C.: Variable metric forward–backward splitting with applications to monotone inclusions in duality. Optimization **63**(9), 1289–1318 (2014). DOI 10.1080/02331934.2012.733883
5. Chouzenoux, E., Pesquet, J.C., Repetti, A.: Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. **162**(1), 107–132 (2014). DOI 10.1007/s10957-013-0465-7
6. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optimization Theory and Applications **158**(2), 460–479 (2013). DOI 10.1007/s10957-012-0245-9
7. Combettes, P.L., Condat, L., Pesquet, J.C., Vu, B.C.: A forward-backward view of some primal-dual optimization methods in image recovery. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 4141–4145 (2014). DOI 10.1109/ICIP.2014.7025841
8. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numerica **25**, 161–319 (2016). DOI 10.1017/S096249291600009X
9. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. **67**(1), 91–108 (2005). DOI 10.1111/j.1467-9868.2005.00490.x
10. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B, Stat. Methodol. **68**(1), 49–67 (2006). DOI 10.1111/j.1467-9868.2005.00532.x
11. Zhang, X., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on bregman iteration. J Sci Comput **46**, 20–46 (2011). DOI 10.1007/s10915-010-9408-8
12. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. Set-Valued and Variational Analysis **20**(2), 307–330 (2012). DOI 10.1007/s11228-011-0191-y
13. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions. Fixed Point Theory and Applications

2016(1), 54 (2016). DOI 10.1186/s13663-016-0543-2

14. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal–dual algorithm. Mathematical Programming **159**(1), 253–287 (2016). DOI 10.1007/s10107-015-0957-3. URL https://doi.org/10.1007/s10107-015-0957-3
15. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Advances in Computational Mathematics **38**(3), 667–681 (2013). DOI 10.1007/s10444-011-9254-8. URL https://doi.org/10.1007/s10444-011-9254-8
16. BoŢ, R.I., Csetnek, E.R., Heinrich, A., Hendrich, C.: On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. Mathematical Programming **150**(2), 251–279 (2015). DOI 10.1007/s10107-014-0766-0. URL https://doi.org/10.1007/s10107-014-0766-0
17. Chambolle, A.: An algorithm for total variation minimization and applications. Journal of Mathematical Imaging and Vision **20**, 89–97 (2004). DOI 10.1023/B:JMIV.0000011325.36760.1e
18. Chambolle, A.: Total variation minimization and a class of binary mrf models. In: Energy Minimization Methods in Computer Vision and Pattern Recognition, *Lecture Notes in Computer Science*, vol. 3757, pp. 136–152 (2005). DOI 10.1007/11585978_10
19. Aujol, J.F.: Some first-order algorithms for total variation based image restoration. J Math Imaging Vis **34**, 307–327 (2009). DOI 10.1007/s10851-009-0149-y
20. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. Image Processing, IEEE Transactions on **18**(11), 2419 –2434 (2009). DOI 10.1109/TIP.2009.2028250
21. Bonettini, S., Loris, I., Porta, F., Prato, M.: Variable metric inexact line-search based methods for nonsmooth optimization. Siam Journal on Optimization **26**(2), 891–921 (2016). DOI 10.1137/15M1019325
22. Bonettini, S., Loris, I., Porta, F., Prato, M., Rebegoldi, S.: On the convergence of a linesearch based proximal-gradient method for nonconvex optimization. Inverse Problems **33**(5), 055,005 (2017). DOI 10.1088/1361-6420/aa5bfd
23. Salzo, S., Villa, S.: Inexact and accelerated proximal point algorithms. Journal of Convex Analysis **19**(4), 1167–1192 (2012)

24. Schmidt, M., Roux, N.L., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11, pp. 1458–1466. Curran Associates Inc., USA (2011)

25. Hu, Y., Chi, E.C., Allen, G.I.: Splitting Methods in Communication, Imaging, Science, and Engineering, chap. ADMM Algorithmic Regularization Paths for Sparse Statistical Machine Learning, pp. 433–460. Springer (2016). DOI 10.1007/978-3-319-41589-5

26. Rose, S., Andersen, M., Sidky, E., Pan, X.: Noise properties of CT images reconstructed by use of constrained total-variation, data-discrepancy minimization. Medical Physics **42**(5), 2690–2698 (2015). DOI 10.1118/1.4914148

27. Rose, S., Andersen, M.S., Sidky, E.Y., Pan, X.: Technical note: Proximal ordered subsets algorithms for TV constrained optimization in CT image reconstruction. Tech. rep., The University of Chicago (2016). ArXiv:1603.08889v1

28. Loris, I., Verhoeven, C.: On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Problems **27**(12), 125,007 (2011). DOI 10.1088/0266-5611/27/12/125007

29. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. Inverse Problems **29**(2), 025,011– (2013). DOI 10.1088/0266-5611/29/2/025011

30. Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. Operations Research Letters **43**(2), 209 – 214 (2015). DOI 10.1016/j.orl.2015.02.001

31. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS book in mathematics. Springer (2011). DOI 10.1007/978-1-4419-9467-7

32. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)

33. Condat, L.: A generic proximal algorithm for convex optimization – application to total variation minimization. IEEE Signal Proc. Letters **21**(8), 1054–1057 (2014). DOI 10.1109/LSP.2014.2322123

34. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J Math Imaging Vis **40**, 120–145 (2011). DOI 10.1007/s10851-010-0251-1

35. Hiriart-Urruty, J.B., Lemarechal, C.: Convex analysis and minimization algorithms. Springer (1993)

36. Chaux, C., Pesquet, J.C., Pustelnik, N.: Nested iterative algorithms for convex constrained image recovery problems. SIAM J. Imaging Sci. **2**(2), 730–762 (2009). DOI 10.1137/080727749

37. Beck, A.: First order methods in optimization. MOS-SIAM Series on Optimization. SIAM (2017)

38. Bonettini, S., Zanella, R., Zanni, L.: A scaled gradient projection method for constrained image deblurring. Inverse Problems **25**(1), 015,002 (2009). DOI 10.1088/0266-5611/25/1/015002

39. Nesterov, Y.E.: A method for solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. Soviet Math. Dokl. **27**, 372–376 (1983)

40. Beck, A., Teboulle, M.: A fast iterative shrinkage-threshold algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**, 183–202 (2009). DOI 10.1137/080716542

41. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". Journal of Optimization Theory and Applications **166**(3), 968–982 (2015). DOI 10.1007/s10957-015-0746-4

42. Chen, J.: Domain decomposition methods and convex optimization with applications to inverse problems. Ph.D. thesis, East China Normal University and Université libre de Bruxelles (2018)