J Math Imaging Vis
DOI 10.1007/s10851-016-0692-2

Acceleration of the PDHGM on Partially Strongly Convex Functions

Tuomo Valkonen[1,2] · **Thomas Pock**[3,4]

Received: 19 April 2016 / Accepted: 23 November 2016

Abstract We propose several variants of the primal–dual method due to Chambolle and Pock. Without requiring full strong convexity of the objective functions, our methods are accelerated on subspaces with strong convexity. This yields mixed rates, $O(1/N^2)$ with respect to initialisation and $O(1/N)$ with respect to the dual sequence, and the residual part of the primal sequence. We demonstrate the efficacy of the proposed methods on image processing problems lacking strong convexity, such as total generalised variation denoising and total variation deblurring.

Keywords Primal–dual · Accelerated · Subspace · Total generalised variation

Mathematics Subject Classification 90C25 · 49M29 · 94A08

✉ Tuomo Valkonen
tuomo.valkonen@iki.fi

Thomas Pock
pock@icg.tugraz.at

[1] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

[2] Department of Mathematical Sciences, University of Liverpool, Liverpool, UK

[3] Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria

[4] Digital Safety and Security Department, AIT Austrian Institute of Technology GmbH, 1220 Vienna, Austria

1 Introduction

Let $G : X \to \overline{\mathbb{R}}$ and $F : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous functionals on Hilbert spaces X and Y, possibly infinite dimensional. Also let $K \in \mathcal{L}(X; Y)$ be a bounded linear operator. We then wish to solve the problem

$$\min_{x \in X} G(x) + F(Kx).$$

This can under mild conditions on F (see, for example, [1,2]) also be written with the help of the convex conjugate F^* in the minimax form

$$\min_{x \in X} \max_{y \in Y} G(x) + \langle Kx, y \rangle - F^*(y).$$

One possibility for the numerical solution of the latter form is the primal–dual algorithm of Chambolle and Pock [3], a type of proximal point or extragradient method, also classified as the 'modified primal–dual hybrid gradient method' or PDHGM by Esser et al. [4]. If either G or F^* is strongly convex, the method can be accelerated to $O(1/N^2)$ convergence rates of the iterates and an ergodic duality gap [3]. But what if we have only partial strong convexity? For example, what if

$$G(x) = G_0(Px)$$

for a projection operator P to a subspace $X_0 \subset X$, and strongly convex $G_0 : X_0 \to \mathbb{R}$? This kind of structure is common in many applications in image processing and data science, as we will more closely review in Sect. 5. Under such *partial strong convexity*, can we obtain a method that would give an accelerated rate of convergence at least for Px?

Published online: 15 December 2016

We provide a partially positive answer: we can obtain mixed rates, $O(1/N^2)$ with respect to initialisation, and $O(1/N)$ with respect to bounds on the 'residual variables' y and $(I - P)x$. In this respect, our results are similar to the 'optimal' algorithm of Chen et al. [5]. Instead of strong convexity, they assume smoothness of G to derive a primal–dual algorithm based on backward–forward steps, instead of the backward–backward steps of [3].

The derivation of our algorithms is based, firstly, on replacing simple step length parameters by a variety of abstract step length operators and, secondly, a type of abstract partial strong monotonicity property

$$\begin{aligned}&\langle \partial G(x') - \partial G(x), \widetilde{T}^{-1}(x' - x)\rangle \\ &\quad \geq \|x' - x\|^2_{\widetilde{T}^{-1,*}\Gamma'} - \text{penalty\_term}, \end{aligned} \tag{1}$$

the full details of which we provide in Sect. 2. Here $\widetilde{T}$ is an auxiliary step length operator. Our factor of strong convexity is a positive semidefinite operator $\Gamma \geq 0$; however, to make our algorithms work, we need to introduce additional artificial strong convexity through another operator Γ', which may not satisfy $0 \leq \Gamma' \leq \Gamma$. This introduces the penalty term in (1). The exact procedure can be seen as a type of smoothing, famously studied by Nesterov [6], and more recently, for instance, by Beck and Teboulle [7]. In these approaches, one computes *a priori* a level of smoothing—comparable to Γ'—needed to achieve prescribed solution quality. One then solves a smoothed problem, which can be done at $O(1/N^2)$ rate. However, to obtain a solution with higher quality than the a priori prescribed one, one needs to solve a new problem from scratch, as the smoothing alters the problem being solved. One can also employ restarting strategies, to take some advantage of the previous solution, see, for example, [8]. Our approach does not depend on restarting and a priori chosen solution qualities: the method will converge to an optimal solution to the original non-smooth problem. Indeed, the introduced additional strong convexity Γ' is controlled automatically.

The 'fast dual proximal gradient method', or FDPG [9], also possesses different type of mixed rates, $O(1/N)$ for the primal, and $O(1/N^2)$ for the dual. This is, however, under standard strong convexity assumptions. Other than that, our work is related to various further developments from the PDHGM, such as variants for nonlinear K [10,11] and non-convex G [12]. The PDHGM has been the basis for inertial methods for monotone inclusions [13] and primal–dual stochastic coordinate descent methods without separability requirements [14]. Finally, the FISTA [15,16] can be seen as a primal-only relative of the PDHGM. Not attempting to do full justice here to the large family of closely related methods, we point to [4,17,18] for further references.

The contributions of our paper are twofold: firstly, to paint a bigger picture of what is possible, we derive a very general version of the PDHGM. This algorithm, useful as a basis for deriving other new algorithms besides ours, is the content of Sect. 2. In this section, we provide an abstract bound on the iterates of the algorithm, later used to derive convergence rates. In Sect. 3, we extend the bound to include an ergodic duality gap under stricter conditions on the acceleration scheme and the step length operators. A by-product of this work is the shortest convergence rate proof for the accelerated PDHGM known to us. Afterwards, in Sect. 4, we derive from the general algorithm two efficient mixed-rate algorithms for problems exhibiting strong convexity only on subspaces. The first one employs the penalty or smoothing ψ on both the primal and the dual. The second one only employs the penalty on the dual. We finish the study with numerical experiments in Sect. 5. The main results of interest for readers wishing to apply our work are Algorithms 3 and 4 along with the respective convergence results, Theorems 4.1 and 4.2.

2 A General Primal–Dual Method

2.1 Notation

To make the notation definite, we denote by $\mathcal{L}(X; Y)$ the space of bounded linear operators between Hilbert spaces X and Y. For $T, S \in \mathcal{L}(X; X)$, the notation $T \geq S$ means that $T - S$ is positive semidefinite; in particular, $T \geq 0$ means that T is positive semidefinite. In this case, we also denote

$$[0, T] := \{\lambda T \mid \lambda \in [0, 1]\}.$$

The identity operator is denoted by I, as is standard.

For $0 \leq M \in \mathcal{L}(X; X)$, which can possibly not be self-adjoint, we employ the notation

$$\langle a, b\rangle_M := \langle Ma, b\rangle, \quad \text{and} \quad \|a\|_M := \sqrt{\langle a, a\rangle_M}. \tag{2}$$

We also use the notation $T^{-1,*} := (T^{-1})^*$.

2.2 Background

As in the introduction, let us be given convex, proper, lower semicontinuous functionals $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ on Hilbert spaces X and Y, as well as a bounded linear operator $K \in \mathcal{L}(X; Y)$. We then wish to solve the minimax problem

$$\min_{x \in X} \max_{y \in Y} \; G(x) + \langle Kx, y\rangle - F^*(y), \tag{P}$$

assuming the existence of a solution $\widehat{u} = (\widehat{x}, \widehat{y})$ satisfying the optimality conditions

$$-K^*\widehat{y} \in \partial G(\widehat{x}), \quad \text{and} \quad K\widehat{x} \in \partial F^*(\widehat{y}). \tag{OC}$$

Such a point always exists if $\lim_{\|x\|\to\infty} G(x)/\|x\| = \infty$ and $\lim_{\|y\|\to\infty} F^*(y)/\|y\| = \infty$, as follows from [2, Proposition VI.1.2 & Proposition VI.2.2]. More generally the existence has to be proved explicitly. In finite dimensions, see, for example, [19] for several sufficient conditions.

The primal–dual method of Chambolle and Pock [3] for the solving (P) consists of iterating

$$x^{i+1} := (I + \tau_i \partial G)^{-1}(x^i - \tau_i K^* y^i), \tag{3a}$$
$$\bar{x}^{i+1} := \omega_i (x^{i+1} - x^i) + x^{i+1}, \tag{3b}$$
$$y^{i+1} := (I + \sigma_{i+1} \partial F^*)^{-1}(y^i + \sigma_{i+1} K \bar{x}^{i+1}). \tag{3c}$$

In the basic version of the algorithm, $\omega_i = 1$, $\tau_i \equiv \tau_0$, and $\sigma_i \equiv \sigma_0$, assuming that the step length parameters satisfy $\tau_0\sigma_0\|K\|^2 < 1$. The method has $O(1/N)$ rate for the ergodic duality gap [3]. If G is strongly convex with factor γ, we may use the acceleration scheme [3]

$$\omega_i := 1/\sqrt{1 + 2\gamma\tau_i}, \quad \tau_{i+1} := \tau_i \omega_i, \quad \text{and} \quad \sigma_{i+1} := \sigma_i/\omega_i, \tag{4}$$

to achieve $O(1/N^2)$ convergence rates of the iterates and an ergodic duality gap, defined in [3]. To motivate our choices later on, observe that σ_0 is never used expect to calculate σ_1. We may therefore equivalently parametrise the algorithm by $\delta = 1 - \|K\|^2\tau_0\sigma_0 > 0$.

We note that the order of the steps in (3) is different from the original ordering in [3]. This is because with the present order, the method (3) may also be written in the proximal point form. This formulation, first observed in [20] and later utilised in [10,11,21], is also what we will use to streamline our analysis. Introducing the general variable splitting notation,

$$u = (x, y),$$

the system (3) then reduces into

$$0 \in H(u^{i+1}) + M_{\text{basic},i}(u^{i+1} - u^i), \tag{5}$$

for the monotone operator

$$H(u) := \begin{pmatrix} \partial G(x) + K^* y \\ \partial F^*(y) - Kx \end{pmatrix}, \tag{6}$$

and the *preconditioning* or *step length operator*

$$M_{\text{basic},i} := \begin{pmatrix} I/\tau_i & -K^* \\ -\omega_i K & I/\sigma_{i+1} \end{pmatrix}. \tag{7}$$

We note that the optimality conditions (OC) can also be encoded as $0 \in H(\widehat{u})$.

2.3 Abstract Partial Monotonicity

Our plan now is to formulate a general version of (3), replacing τ_i and σ_i by operators $T_i \in \mathcal{L}(X; X)$ and $\Sigma_i \in \mathcal{L}(Y; Y)$. In fact, we will need two additional operators $\widetilde{T}_i \in \mathcal{L}(X; X)$ and $\hat{T}_i \in \mathcal{L}(Y; Y)$ to help communicate change in T_i to Σ_i. They replace ω_i in (3b) and (7), operating as $\hat{T}_{i+1} K \widetilde{T}_i^{-1} \approx \omega_i K$ from both sides of K. The role of $\widetilde{T}_i$ is to split the original primal step length τ_i in the space X into the two parts T_i and $\widetilde{T}_i$ with potentially different rates. The role of $\hat{T}_i$ is to transfer $\widetilde{T}_i$ into the space Y, to eventually control the dual step length Σ_i. In the basic algorithm (3), we would simply have $\widetilde{T}_i = T_i = \tau_i I \in \mathcal{L}(X; X)$, and $\hat{T}_i = \tau_i I \in \mathcal{L}(Y; Y)$ for the scalar τ_i.

To start the algorithm derivation, we now formulate abstract forms of partially strong monotonicity. As a first step, we take subsets of invertible operators

$$\widetilde{\mathcal{T}} \subset \mathcal{L}(X; X), \quad \text{and} \quad \hat{\mathcal{T}} \subset \mathcal{L}(Y; Y),$$

as well as subsets of positive semidefinite operators

$$0 \le \widetilde{\mathcal{K}} \subset \mathcal{L}(X; X), \quad \text{and} \quad 0 \le \hat{\mathcal{K}} \subset \mathcal{L}(Y; Y).$$

We assume $\widetilde{\mathcal{T}}$ and $\hat{\mathcal{T}}$ closed with respect to composition: $\widetilde{T}_1\widetilde{T}_2 \in \widetilde{\mathcal{T}}$ for $\widetilde{T}_1, \widetilde{T}_2 \in \widetilde{\mathcal{T}}$.

We use the sets $\widetilde{\mathcal{K}}$ and $\hat{\mathcal{K}}$ as follows. We suppose that ∂G is *partially strongly* $(\psi, \widetilde{\mathcal{T}}, \widetilde{\mathcal{K}})$-*monotone*, meaning that for all $x, x' \in X$, $\widetilde{T} \in \widetilde{\mathcal{T}}$, $\Gamma' \in [0, \Gamma] + \widetilde{\mathcal{K}}$ holds

$$\begin{aligned} &\langle \partial G(x') - \partial G(x), \widetilde{T}^{-1}(x' - x)\rangle \\ &\quad \ge \|x' - x\|^2_{\widetilde{T}^{-1,*}\Gamma'} - \psi_{\widetilde{T}^{-1,*}(\Gamma'-\Gamma)}(x' - x), \end{aligned} \tag{G-PM}$$

for some family of functionals $\{\psi_T : X \to \mathbb{R}\}$, and a linear operator $0 \le \Gamma \in \mathcal{L}(X; X)$ which models partial strong monotonicity. The inequality in (G-PM), and all such set inequalities in the remainder of this paper, is understood to hold for all elements of the sets $\partial G(x')$ and $\partial G(x)$. The operator $\widetilde{T} \in \widetilde{\mathcal{T}}$ acts as a *testing operator*, and the operator $\Gamma' \in \widetilde{\mathcal{K}}$ as *introduced strong monotonicity*. The functional $\psi_{\widetilde{T}^{-1,*}(\Gamma'-\Gamma)}$ is a *penalty* corresponding to the test and the introduced strong monotonicity. The role of testing will become more apparent in Sect. 2.4.

Similarly to (G-PM), we assume that ∂F^* is $(\phi, \hat{\mathcal{T}}, \hat{\mathcal{K}})$-*monotone* with respect to $\hat{\mathcal{T}}$ in the sense that for all $y, y' \in Y$, $\hat{T} \in \hat{\mathcal{T}}$, $R \in \hat{\mathcal{K}}$ holds

$$\begin{aligned} &\langle \partial F^*(y') - \partial F^*(y), \hat{T}^{-1,*}(y' - y)\rangle \\ &\quad \ge \|y' - y\|^2_{\hat{T}^{-1}R} - \phi_{\hat{T}^{-1}R}(y' - y), \end{aligned} \tag{F*-PM}$$

for some family of functionals $\{\phi_T : Y \to \mathbb{R}\}$. Again, the inequality in (F*-PM) is understood to hold for all elements of the sets $\partial F^*(y')$ and $\partial F^*(y)$.

In our general analysis, we do not set any conditions on ψ and ϕ, as their role is simply symbolic transfer of dissatisfaction of strong monotonicity into a penalty in our abstract convergence results.

Let us next look at a few examples on how (G-PM) or (F*-PM) might be satisfied. First we have the very well-behaved case of quadratic functions.

Example 2.1 $G(x) = \|f - Ax\|^2/2$ satisfies (G-PM) with $\Gamma = A^*A$, $\widetilde{\mathcal{K}} = \{0\}$, and $\psi \equiv 0$ for any invertible $\widetilde{T}$. Indeed, G is differentiable with $\langle \nabla G(x') - \nabla G(x), \widetilde{T}^{-1}(x' - x)\rangle = \langle A^*A(x' - x), \widetilde{T}^{-1}(x' - x)\rangle = \|x' - x\|^2_{\widetilde{T}^{-1,*}\Gamma}$.

The next lemma demonstrates what can be done when all the parameters are scalar. It naturally extends to functions of the form $G(x_1, x_2) = G(x_1) + G(x_2)$ with corresponding product form parameters.

Lemma 2.1 *Let $G : X \to \overline{\mathbb{R}}$ be convex and proper with* dom G *bounded. Then,*

$$G(x') - G(x) \geq \langle \partial G(x), x' - x\rangle + \frac{\gamma}{2}\left(\|x' - x\|^2 - C_\psi\right), \tag{8}$$

for some constant $C_\psi \geq 0$, every $\gamma \geq 0$, and $x, x' \in X$.

Proof We denote $A := \operatorname{dom} G$. If $x' \notin A$, we have $G(x') = \infty$, so (8) holds irrespective of γ and C. If $x \notin A$, we have $\partial G(x) = \emptyset$, so (8) again holds. We may therefore compute the constants based on $x, x' \in A$. Now, there is a constant M such that $\sup_{x\in A} \|x\| \leq M$. Then, $\|x' - x\| \leq 2M$. Thus, if we pick $C = 4M^2$, then $(\gamma/2)(\|x' - x\|^2 - C) \leq 0$ for every $\gamma \geq 0$ and $x, x' \in A$. By the convexity of G, (8) holds. □

Example 2.2 An indicator function ι_A of a convex bounded set A satisfies the conditions of Lemma 2.1. This is generally what we will use and need.

2.4 A General Algorithm and the Idea of Testing

The only change we make to the proximal point formulation (5) of the method (3) is to replace the basic step length or preconditioning operator $M_{\text{basic},i}$ by the operator

$$M_i := \begin{pmatrix} T_i^{-1} & -K^* \\ -\hat{T}_{i+1}K\widetilde{T}_i^{-1} & \Sigma_{i+1}^{-1}\end{pmatrix}. \tag{10}$$

As we have remarked, the operators $\hat{T}_{i+1}$ and $\widetilde{T}_i$ play the role of ω_i, acting from both sides of K. Our proposed algorithm can thus be characterised as solving on each iteration $i \in \mathbb{N}$ for the next iterate u^{i+1} the preconditioned proximal point problem

$$0 \in H(u^{i+1}) + M_i(u^{i+1} - u^i). \tag{PP}$$

To study the convergence properties of (PP), we define the *testing operator*

$$S_i := \begin{pmatrix} \widetilde{T}_i^{-1,*} & 0 \\ 0 & \hat{T}_{i+1}^{-1}\end{pmatrix}. \tag{11}$$

It will turn out that multiplying or 'testing' (PP) by this operator will allow us to derive convergence rates. The testing of (PP) by S_i is why we introduced testing into the monotonicity conditions (G-PM) and (F*-PM). If we only tested (PP) with $S_i = I$, we could at most obtain ergodic convergence of the duality gap for the unaccelerated method. But by testing with something appropriate and faster increasing, such as (11), we are able to extract better convergence rates from (PP).

We also set

$$\bar{\Gamma}_i = \begin{pmatrix} 2\Gamma_i & \widetilde{T}_i^*(K\widetilde{T}_i^{-1} - \hat{T}_{i+1}^{-1}K)^* \\ \hat{T}_{i+1}(K\widetilde{T}_i^{-1} - \hat{T}_{i+1}^{-1}K) & 2R_{i+1}\end{pmatrix},$$

for some $\Gamma_i \in [0, \Gamma] + \widetilde{\mathcal{K}}$ and $R_{i+1} \in \hat{\mathcal{K}}$. We will see in Sect. 2.6 that $\bar{\Gamma}_i$ is a factor of partial strong monotonicity for H with respect to testing by S_i. With this, taking a fixed $\delta > 0$, the properties

$$S_i(M_i + \bar{\Gamma}_i) \geq S_{i+1}M_{i+1}, \quad \text{and} \tag{C1}$$

Algorithm 1 Primal–dual algorithm with partial acceleration

Require: F^* and G satisfying (G-PM) and (F*-PM) for some sets and spaces $\widetilde{\mathcal{K}}, \hat{\mathcal{K}}, \widetilde{\mathcal{T}}, \hat{\mathcal{T}}$, and $0 \leq \Gamma \in \mathcal{L}(X; X)$. Initial invertible $T_0 \in \mathcal{L}(X; X)$, $\widetilde{T}_0 \in \widetilde{\mathcal{T}}$, $\hat{T}_1 \in \hat{\mathcal{T}}$, and $\Sigma_1 \in \mathcal{L}(Y; Y)$, as well as $\delta \in (0, 1)$, satisfying for $j = 0$ the condition

$$S_jM_j \geq \delta\begin{pmatrix} \widetilde{T}_j^{-1,*}T_j^{-1} & 0 \\ 0 & 0\end{pmatrix}. \tag{9}$$

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **repeat**
3: Perform the updates

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_iK^*y^i),$$
$$\bar{w}^{i+1} := \hat{T}_{i+1}K\widetilde{T}_i^{-1}(x^{i+1} - x^i) + Kx^{i+1},$$
$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \Sigma_{i+1}\bar{w}^{i+1}).$$

4: Find invertible $T_{i+1} \in \mathcal{L}(X; X)$, $\widetilde{T}_{i+1} \in \widetilde{\mathcal{T}}$, $\hat{T}_{i+2} \in \hat{\mathcal{T}}$, and $\Sigma_{i+2} \in \mathcal{L}(Y; Y)$ satisfying (9) with $j = i + 1$, as well as the condition

$$S_i(M_i + \bar{\Gamma}_i) \geq S_{i+1}M_{i+1}$$

for some $0 \leq R_{i+1} \in \hat{\mathcal{K}}$ and $\Gamma_i \in [0, \Gamma] + \widetilde{\mathcal{K}}$.
5: **until** a stopping criterion is fulfilled.

$$S_i M_i \geq \delta \begin{pmatrix} \widetilde{T}_i^{-1,*} T_i^{-1} & 0 \\ 0 & 0 \end{pmatrix} \geq 0, \tag{C2}$$

will turn out to be the crucial defining properties for the convergence rates of the iteration (PP). The method resulting from the combination of (PP), (C1), and (C2) can also be expressed as Algorithm 1. The main steps in developing practical algorithms based on Algorithm 1 will be in the choice of the various step length operators. This will be the content of Sects. 3 and 4. Before this, we expand the conditions (C1) and (C2) to see how they might be satisfied and study abstract convergence results.

2.5 A Simplified Condition

We expand

$$S_i M_i = \begin{pmatrix} \widetilde{T}_i^{-1,*} T_i^{-1} & -\widetilde{T}_i^{-1,*} K^* \\ -K \widetilde{T}_i^{-1} & \hat{T}_{i+1}^{-1} \Sigma_{i+1}^{-1} \end{pmatrix}, \tag{12}$$

as well as

$$S_i \bar{\Gamma}_i = \begin{pmatrix} 2\widetilde{T}_i^{-1,*} \Gamma_i & \widetilde{T}_i^{-1,*} K^* - K^* \hat{T}_{i+1}^{-1,*} \\ K \widetilde{T}_i^{-1} - \hat{T}_{i+1}^{-1} K & 2\hat{T}_{i+1}^{-1} R_{i+1} \end{pmatrix}, \tag{13}$$

and

$$S_i (M_i + \bar{\Gamma}_i) = \begin{pmatrix} \widetilde{T}_i^{-1,*} (T_i^{-1} + 2\Gamma_i) & -K^* \hat{T}_{i+1}^{-1,*} \\ -\hat{T}_{i+1}^{-1} K & \hat{T}_{i+1}^{-1} (\Sigma_{i+1}^{-1} + 2R_{i+1}) \end{pmatrix}.$$

We observe that if $S, T \in \mathcal{L}(X; Y)$, then for arbitrary invertible $Z \in \mathcal{L}(Y; Y)$ a type of Cauchy (or Young) inequality holds, namely

$$\begin{aligned} \begin{pmatrix} 0 & T^* S \\ S^* T & 0 \end{pmatrix} &= \begin{pmatrix} 0 & T^* Z^* Z^{-1,*} S \\ S^* Z^{-1} Z T & 0 \end{pmatrix} \\ &\leq \begin{pmatrix} T^* Z^* Z T & 0 \\ 0 & S^* Z^{-1} Z^{-1,*} S \end{pmatrix}. \end{aligned} \tag{14}$$

The inequality here can be verified using the basic Cauchy inequality $2\langle x, y\rangle \leq \|x\|^2 + \|y\|^2$. Applying (14) in (12), we see that (C2) is satisfied when

$$\begin{aligned} &\hat{T}_{i+1}^{-1} \Sigma_{i+1}^{-1} \geq K Z_i^{-1} Z_i^{-1,*} K^*, \quad \text{and} \\ &(1-\delta) \widetilde{T}_i^{-1,*} T_i^{-1} \geq \widetilde{T}_i^{-1,*} Z_i^* Z_i \widetilde{T}^{-1}, \end{aligned} \tag{15}$$

for some invertible $Z_i \in \mathcal{L}(X; X)$. The second condition in (15) is satisfied as an equality if

$$Z_i^* Z_i = (1-\delta) T_i^{-1} \widetilde{T}_i. \tag{16}$$

By the spectral theorem for self-adjoint operators on Hilbert spaces (e.g. [22, Chapter 12]), we can make the choice (16) if

$$\begin{aligned} &T_i^{-1} \widetilde{T}_i \in \mathcal{Q} \\ &:= \{A \in \mathcal{L}(X; X) \mid A \text{ is self-adjoint and positive definite}\}. \end{aligned}$$

Equivalently, by the same spectral theorem, $\widetilde{T}_i^{-1} T_i \in \mathcal{Q}$. Therefore, we see from (15) that (C2) holds when

$$\widetilde{T}_i^{-1} T_i \in \mathcal{Q} \quad \text{and} \quad \hat{T}_{i+1}^{-1} \Sigma_{i+1}^{-1} \geq \frac{1}{1-\delta} K \widetilde{T}_i^{-1} T_i K^*. \tag{C2$'$}$$

Also, (C1) can be rewritten as

$$\begin{pmatrix} \widetilde{T}_i^{-1,*} (T_i^{-1} + 2\Gamma_i) - \widetilde{T}_{i+1}^{-1,*} T_{i+1}^{-1} & \widetilde{T}_{i+1}^{-1,*} K^* - K^* \hat{T}_{i+1}^{-1,*} \\ K \widetilde{T}_{i+1}^{-1} - \hat{T}_{i+1}^{-1} K & \hat{T}_{i+1}^{-1} (\Sigma_{i+1}^{-1} + 2R_{i+1}) - \hat{T}_{i+2}^{-1} \Sigma_{i+2}^{-1} \end{pmatrix} \geq 0. \tag{C1$'$}$$

2.6 Basic Convergence Result

Our main result on Algorithm 1 is the following theorem, providing some general convergence estimates. It is, however, important to note that the theorem does not yet directly prove convergence, as its estimates depend on the rate of decrease in $T_N \widetilde{T}_N^*$, as well as the rate of increase in the penalty sum $\sum_{i=0}^{N-1} D_{i+1}$ coming from the dissatisfaction of strong convexity. Deriving these rates in special cases will be the topic of Sect. 4.

Theorem 2.1 *Let us be given $K \in \mathcal{L}(X; Y)$, and convex, proper, lower semicontinuous functionals $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ on Hilbert spaces X and Y, satisfying* (G-PM) *and* (F*-PM). *Pick $\delta \in (0, 1)$, and suppose* (C1) *and* (C2) *are satisfied for each $i \in \mathbb{N}$ for some invertible $T_i \in \mathcal{L}(X; X)$, $\widetilde{T}_i \in \widetilde{\mathcal{T}}$, $\hat{T}_{i+1} \in \hat{\mathcal{T}}$, and $\Sigma_{i+1} \in \mathcal{L}(Y; Y)$, as well as $\Gamma_i \in [0, \Gamma] + \widetilde{\mathcal{K}}$ and $R_{i+1} \in \hat{\mathcal{K}}$. Suppose that $\widetilde{T}_i^{-1,*} T_i^{-1}$ and $\hat{T}_{i+1}^{-1} \Sigma_{i+1}^{-1}$ are self-adjoint. Let $\widehat{u} = (\widehat{x}, \widehat{y})$ satisfy* (OC). *Then, the iterates of Algorithm 1 satisfy*

$$\frac{\delta}{2} \|x^N - \widehat{x}\|^2_{\widetilde{T}_N^{-1,*} T_N^{-1}} \leq C_0 + \sum_{i=0}^{N-1} \widetilde{D}_{i+1}, \quad (N \geq 1), \tag{17}$$

for

$$\begin{aligned} \widetilde{D}_{i+1} &:= \psi_{\widetilde{T}_i^{-1,*}(\Gamma_i - \Gamma)}(x^{i+1} - \widehat{x}) \\ &+ \phi_{\hat{T}_{i+1}^{-1} R_{i+1}}(y^{i+1} - \widehat{y}), \quad \text{and} \quad C_0 := \frac{1}{2} \|u^0 - \widehat{u}\|^2_{S_0 M_0}. \end{aligned} \tag{18}$$

Remark 2.1 The term $\widetilde{D}_{i+1}$, coming from the dissatisfaction of strong convexity, penalises the basic convergence, which is on the right-hand side of (17) presented by the constant C_0. If $T_N\widetilde{T}_N$ is of the order $O(1/N^2)$, at least on a subspace, and we can bound the penalty $\widetilde{D}_{i+1} \leq C$ for some constant C, then we clearly obtain mixed $O(1/N^2) + O(1/N)$ convergence rates on the subspace. If we can assume that $\widetilde{D}_{i+1}$ actually converges to zero at some rate, then it will even be possible to obtain improved convergence rates. Since typically $\widetilde{T}_i, \hat{T}_{i+1} \searrow 0$ reduce to scalar factors within $\widetilde{D}_{i+1}$, this would require prior knowledge of the rates of convergence $x^i \to \widehat{x}$ and $y^i \to \widehat{y}$. Boundedness of the iterates $\{(x^i, y^i)\}_{i=0}^{\infty}$, we can, however, usually ensure.

Proof Since $0 \in H(\widehat{u})$, we have

$$\begin{aligned}\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \subset \langle H(u^{i+1}) \\ - H(\widehat{u}), S_i^*(u^{i+1} - \widehat{u})\rangle.\end{aligned}$$

Recalling the definition of S_i from (11), and of H from (6), it follows

$$\begin{aligned}&\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \subset \langle \partial G(x^{i+1}) \\ &- \partial G(\widehat{x}), \widetilde{T}_i^{-1}(x^{i+1} - \widehat{x})\rangle \\ &+ \langle \partial F^*(y^{i+1}) - \partial F^*(\widehat{y}), \hat{T}_{i+1}^{-1,*}(y^{i+1} - \widehat{y})\rangle \\ &+ \langle K^*(y^{i+1} - \widehat{y}), \widetilde{T}_i^{-1}(x^{i+1} - \widehat{x})\rangle \\ &- \langle K(x^{i+1} - \widehat{x}), \hat{T}_{i+1}^{-1,*}(y^{i+1} - \widehat{y})\rangle.\end{aligned}$$

An application of (G-PM) and (F*-PM) consequently gives

$$\begin{aligned}&\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \\ &\geq \|x^{i+1} - \widehat{x}\|^2_{\widetilde{T}_i^{-1,*}\Gamma_i} + \|y^{i+1} - \widehat{y}\|^2_{\hat{T}_{i+i}^{-1}R_{i+1}} \\ &\quad - \phi_{\hat{T}_{i+1}^{-1}R_{i+1}}(y^{i+1} - \widehat{y}) - \psi_{\widetilde{T}_i^{-1,*}(\Gamma_i - \Gamma)}(x^{i+1} - \widehat{x}) \\ &\quad + \langle K\widetilde{T}_i^{-1}(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle \\ &\quad - \langle \hat{T}_{i+1}^{-1}K(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle.\end{aligned}$$

Using the expression (13) for $S_i\bar{\Gamma}_i$, and (18) for $\widetilde{D}_{i+1}$, we thus deduce

$$\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \geq \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_i\bar{\Gamma}_i} - \widetilde{D}_{i+1}. \tag{19}$$

For arbitrary self-adjoint $M \in \mathcal{L}(X \times Y; X \times Y)$, we calculate

$$\begin{aligned}\langle u^{i+1} - u^i, u^{i+1} - \widehat{u}\rangle_M = \frac{1}{2}\|u^{i+1} - u^i\|^2_M \\ - \frac{1}{2}\|u^i - \widehat{u}\|^2_M + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_M.\end{aligned} \tag{20}$$

We observe that S_iM_i in (12) is self-adjoint as we have assumed that $\widetilde{T}_i^{-1,*}T_i^{-1}$ and $\hat{T}_{i+1}^{-1}\Sigma_{i+1}^{-1}$ are self-adjoint. In consequence, using (20) we obtain

$$\begin{aligned}\langle M_i(u^i - u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle = -\frac{1}{2}\|u^{i+1} - u^i\|^2_{S_iM_i} \\ + \frac{1}{2}\|u^i - \widehat{u}\|^2_{S_iM_i} - \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_iM_i}.\end{aligned}$$

Using (C1) to estimate $\frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_iM_i}$ and (C2) to eliminate $\frac{1}{2}\|u^{i+1} - u^i\|^2_{S_iM_i}$ yields

$$\begin{aligned}\langle M_i(u^i - u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \leq \frac{1}{2}\|u^i - \widehat{u}\|^2_{S_iM_i} \\ - \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_i\bar{\Gamma}_i}.\end{aligned} \tag{21}$$

Combining (19) and (21) through (PP), we thus obtain

$$\frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_{i+1}M_{i+1}} \leq \frac{1}{2}\|u^i - \widehat{u}\|^2_{S_iM_i} + \widetilde{D}_{i+1}. \tag{22}$$

Summing (22) over $i = 1, \ldots, N-1$, and applying (C2) to estimate S_NM_N from below, we obtain (17). □

3 Scalar Off-diagonal Updates and the Ergodic Duality Gap

One relatively easy way to satisfy (G-PM), (F*-PM), (C1) and (C2) is to take the 'off-diagonal' step length operators $\hat{T}_i$ and $\widetilde{T}_i$ as equal scalars. Another good starting point would be to choose $\widetilde{T}_i = T_i$. We, however, do not explore this route in the present work. Instead, we now specialise Theorem 2.1 to the scalar case. We then explore ways to add estimates of the ergodic duality gap into (17). While this would be possible in the general framework through convexity notions analogous to (G-PM) and (F*-PM), the resulting gap would not be particularly meaningful. We therefore concentrate on the scalar off-diagonal updates to derive estimates on the ergodic duality gap.

3.1 Scalar Specialisation of Algorithm 1

We take both $\widetilde{T}_i = \widetilde{\tau}_i I$, and $\hat{T}_i = \hat{\tau}_i I$ for some $\widetilde{\tau}_i > 0$. With

$$\widetilde{\omega}_i := \widetilde{\tau}_{i+1}/\widetilde{\tau}_i,$$

the condition (C2′) then becomes

$$T_i \in \mathcal{Q}, \quad \text{and} \quad \Sigma_{i+1}^{-1} \geq \widetilde{\omega}_i(1-\delta)^{-1}KT_iK^*. \tag{C2″}$$

The off-diagonal terms cancelling out (C1′) on the other hand become

Algorithm 2 Primal–dual algorithm with partial acceleration—partially scalar

Require: F^* and G satisfying (G-pm) and (F*-pm) for some sets $\widetilde{\mathcal{K}}$, $\hat{\mathcal{K}}$, and $0 \le \Gamma \in \mathcal{L}(X;X)$. A choice of $\delta \in (0,1)$. Initial invertible step length operators $T_0 \in \mathcal{Q}$ and $\Sigma_0 \in \mathcal{L}(Y;Y)$, as well as step length parameter $\widetilde{\tau}_0 > 0$.

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **repeat**
3: Find $\widetilde{\omega}_i > 0$, $\Omega_i \in \mathcal{L}(X;X)$, and $\Gamma_i \in [0,\Gamma] + \widetilde{\mathcal{K}}$ satisfying

$$\widetilde{\omega}_i(I + 2\Gamma_i T_i)\Omega_i \ge I, \quad \text{and} \quad T_i\Omega_i \in \mathcal{Q}. \tag{23a}$$

4: Set

$$T_{i+1} := T_i\Omega_i, \quad \text{and} \quad \widetilde{\tau}_{i+1} := \widetilde{\tau}_i\widetilde{\omega}_i. \tag{23b}$$

5: Find $\Sigma_{i+1} \in \mathcal{L}(Y;Y)$ and $R_i \in \hat{\mathcal{K}}$ satisfying

$$\Sigma_i^{-1} + 2R_i \ge \widetilde{\omega}_i^{-1}\Sigma_{i+1}^{-1} \ge (1-\delta)^{-1}KT_iK^*. \tag{23c}$$

6: Perform the updates

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_iK^*y^i),$$
$$\bar{x}^{i+1} := \widetilde{\omega}_i(x^{i+1} - x^i) + x^{i+1},$$
$$y^{i+1} := (I + \Sigma_{i+1}\partial F^*)^{-1}(y^i + \Sigma_{i+1}K\bar{x}^{i+1}).$$

7: **until** a stopping criterion is fulfilled.

$$\widetilde{\tau}_i^{-1}(I + 2\Gamma_i T_i)T_i^{-1} \ge \widetilde{\tau}_{i+1}^{-1}T_{i+1}^{-1}, \quad \text{and} \tag{C1$''$}$$
$$\widetilde{\tau}_{i+1}^{-1}(\Sigma_{i+1}^{-1} + 2R_{i+1}) \ge \widetilde{\tau}_{i+2}^{-1}\Sigma_{i+2}^{-1}.$$

Observe also that M_i is under this setup self-adjoint if T_i and Σ_{i+1} are.

For simplicity, we now assume ϕ and ψ to satisfy the identities

$$\psi_T(-x) = \psi_T(x), \quad \text{and}$$
$$\psi_{\alpha T}(x) = \alpha\psi_T(x), \quad (x \in X;\ 0 < \alpha \in \mathbb{R}). \tag{24}$$

The monotonicity conditions (G-PM) and (F*-PM) then simplify into

$$\langle \partial G(x') - \partial G(x), x' - x\rangle \ge \|x' - x\|^2_{\Gamma'} - \psi_{\Gamma'-\Gamma}(x' - x), \tag{G-pm}$$

holding for all $x, x' \in X$, and $\Gamma' \in [0,\Gamma] + \widetilde{\mathcal{K}}$; and

$$\langle \partial F^*(y') - \partial F^*(y), y' - y\rangle \ge \|y' - y\|^2_R - \phi_R(y' - y), \tag{F*-pm}$$

holding for all $y, y' \in Y$, and $R \in \hat{\mathcal{K}}$.

We have thus converted the main conditions (C2), (C1), (G-PM), and (F*-PM) of Theorem 2.1 into the respective conditions (C2$''$), (C1$''$), (G-pm), and (F*-pm). Rewriting (C1$''$) in terms of $0 < \Omega_i \in \mathcal{L}(X;X)$ and $\widetilde{\omega}_i > 0$ satisfying

$$T_{i+1} = T_i\Omega_i \quad \text{and} \quad \widetilde{\tau}_{i+1} = \widetilde{\tau}_i\widetilde{\omega}_i,$$

we reorganise (C1$''$) and (C2$''$) into the parameter update rules (23) of Algorithm 2. For ease of expression, we introduce there Σ_0 and R_0 as dummy variables that are not used anywhere else. Equating $\bar{w}^{i+1} = K\bar{x}^{i+1}$, we observe that Algorithm 2 is an instance of Algorithm 1.

Example 3.1 (The method of Chambolle and Pock) Let G be strongly convex with factor $\gamma \ge 0$. We take $T_i = \tau_i I$, $\widetilde{T}_i = \tau_i I$, $\hat{T}_i = \tau_i I$, and $\Sigma_{i+1} = \sigma_{i+1}I$ for some scalars $\tau_i, \sigma_{i+1} > 0$. The conditions (G-pm) and (F*-pm) then hold with $\psi \equiv 0$ and $\phi \equiv 0$, while (C2$''$) and (C1$''$) reduce with $R_{i+1} = 0$, $\Gamma_i = \gamma I$, $\Omega_i = \omega_i I$, and $\widetilde{\omega}_i = \omega_i$ into

$$\omega_i^2(1 + 2\gamma\tau_i) \ge 1, \quad \text{and}$$
$$(1-\delta)/\|K\|^2 \ge \tau_{i+2}\sigma_{i+2} \ge \tau_{i+1}\sigma_{i+1}.$$

Updating σ_{i+1} such that the last inequality holds as an equality, we recover the accelerated PDHGM (3)+(4). If $\gamma = 0$, we recover the unaccelerated PDHGM.

3.2 The Ergodic Duality Gap and Convergence

To study the convergence of an ergodic duality gap, we now introduce convexity notions analogous to (G-pm) and (F*-pm). Namely, we assume

$$G(x') - G(x) \ge \langle \partial G(x), x' - x\rangle + \frac{1}{2}\|x' - x\|^2_{\Gamma'} - \frac{1}{2}\psi_{\Gamma'-\Gamma}(x' - x), \tag{G-pc}$$

to hold for all $x, x' \in X$ and $\Gamma' \in [0,\Gamma] + \widetilde{\mathcal{K}}$ and

$$F^*(y') - F^*(y) \ge \langle \partial F^*(y), y' - y\rangle + \frac{1}{2}\|y' - y\|^2_R - \frac{1}{2}\psi_R(y' - y), \tag{F*-pc}$$

to hold for all $y, y' \in Y$ and $R \in \hat{\mathcal{K}}$. Clearly these imply (G-pm) and (F*-pm).

To define an ergodic duality gap, we set

$$\widetilde{q}_N := \sum_{i=0}^{N-1}\widetilde{\tau}_i^{-1}, \quad \text{and} \quad \hat{q}_N := \sum_{i=0}^{N-1}\widetilde{\tau}_{i+1}^{-1}, \tag{25}$$

and define the weighted averages

$$x_N := \widetilde{q}_N^{-1}\sum_{i=0}^{N-1}\widetilde{\tau}_i^{-1}x^{i+1}, \quad \text{and} \quad y_N := \hat{q}_N^{-1}\sum_{i=0}^{N-1}\widetilde{\tau}_{i+1}^{-1}y^{i+1}.$$

With these, the ergodic duality gap at iteration N is defined as the duality gap for (x_N, Y_N), namely

$$\mathcal{G}^N := \big(G(x_N) + \langle \widehat{y}, Kx_N\rangle - F^*(\widehat{y})\big) - \big(G(\widehat{x}) + \langle y_N, K\widehat{x}\rangle - F^*(y_N)\big),$$

and we have the following convergence result.

Theorem 3.1 *Let us be given $K \in \mathcal{L}(X;Y)$, and convex, proper, lower semicontinuous functionals $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ on Hilbert spaces X and Y, satisfying* (G-pc) *and* (F*-pc) *for some sets $\widetilde{\mathcal{K}}$, $\hat{\mathcal{K}}$, and $0 \le \Gamma \in \mathcal{L}(X;X)$. Pick $\delta \in (0,1)$, and suppose* (C2″) *and* (C1″) *are satisfied for each $i \in \mathbb{N}$ for some invertible self-adjoint $T_i \in \mathcal{Q}$, $\Sigma_i \in \mathcal{L}(Y;Y)$,*

$$0 < \widetilde{\tau}_i \le \widetilde{\tau}_0, \tag{C3″}$$

as well as $\Gamma_i \in \lambda([0,\Gamma] + \widetilde{\mathcal{K}})$ and $R_i \in \lambda\hat{\mathcal{K}}$ for $\lambda = 1/2$. Let $\widehat{u} = (\widehat{x}, \widehat{y})$ satisfy (OC)*. Then, the iterates of Algorithm 2 satisfy*

$$\frac{\delta}{2}\|x^N - \widehat{x}\|^2_{\widetilde{\tau}_N^{-1}T_N^{-1}} + \widetilde{q}_N\mathcal{G}^N \le C_0 + \sum_{i=0}^{N-1} D_{i+1}. \tag{26}$$

Here C_0 is as in (18)*, and*

$$D_{i+1} := \widetilde{\tau}_i^{-1}\psi_{\Gamma_i-\lambda\Gamma}(x^{i+1} - \widehat{x}) + \widetilde{\tau}_{i+1}^{-1}\phi_{R_{i+1}}(y^{i+1} - \widehat{y}). \tag{27}$$

If only (G-pm) *and* (F*-pm) *hold instead of* (G-pc) *and* (F*-pc)*, and we take $\lambda = 1$, then* (26) *holds with the modification $\mathcal{G}^N := 0$.*

Remark 3.1 For convergence of the gap, we must accelerate less (factor $1/2$ on Γ_i).

Example 3.2 (*No acceleration*) Consider Example 3.1, where $\psi \equiv 0$ and $\phi \equiv 0$. If $\gamma = 0$, we get ergodic convergence of the duality gap at rate $O(1/N)$. Indeed, we are in the scalar step setting, with $\widetilde{\tau}_j = \widetilde{\tau}_j = \tau_0$. Thus, presently $\widetilde{q}_N = N\tau_0$.

Example 3.3 (*Full acceleration*) With $\gamma > 0$ in Example 3.1, we know from [3, Corollary 1] that

$$\lim_{N\to\infty} N\tau_N\gamma = 1. \tag{28}$$

Thus, $\widetilde{q}_N$ is of the order $\Omega(N^2)$, while $\widetilde{\tau}_N T_N = \tau_N^2 I$ is of the order $O(1/N^2)$. Therefore, (26) shows $O(1/N^2)$ convergence of the squared distance to solution. For $O(1/N^2)$ convergence of the ergodic duality gap, we need to slow down (4) to $\omega_i = 1/\sqrt{1+\gamma\tau_i}$.

Remark 3.2 The result (28) can be improved to estimate $\tau_N \le C_\tau/N$ without a qualifier $N \ge N_0$. Indeed, from [3, Lemma 2] we know the following for the rule $\omega_i = 1/\sqrt{1+2\gamma\tau_i}$: given $\lambda > 0$ and $N \ge 0$ with $\gamma\tau_N \le \lambda$, for any $\ell \ge 0$ holds

$$\frac{1}{\gamma\tau_N} + \frac{\ell}{1+\lambda} \le \frac{1}{\gamma\tau_{N+\ell}} \le \frac{1}{\gamma\tau_N} + \ell.$$

If we pick $N = 0$ and $\lambda = \gamma\tau_0$, this says

$$\frac{1}{\gamma\tau_0} + \frac{\ell}{1+\gamma\tau_0} \le \frac{1}{\gamma\tau_\ell} \le \frac{1}{\gamma\tau_0} + \ell.$$

The first inequality gives $\tau_\ell \le (1+\gamma\tau_0)/(\tau_0^{-1} + \gamma\ell) \le (\gamma^{-1} + \tau_0)/\ell$.

Therefore, $\tau_N \le C_\tau/N$ for $C_\tau := \gamma^{-1} + \tau_0$. Moreover, the second inequality gives $\tau_N^{-1} \le \tau_0^{-1} + \gamma N$.

Proof (*Theorem 3.1*) The non-gap estimate in the last paragraph of the theorem statement, where $\lambda = 1$, we modify $\mathcal{G}_N := 0$, is a direct consequence of Theorem 2.1. We therefore concentrate on the estimate that includes the gap, and fix $\lambda = 1/2$. We begin by expanding

$$\begin{aligned}
&\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \\
&\quad = \widetilde{\tau}_i^{-1}\langle \partial G(x^{i+1}), x^{i+1} - \widehat{x}\rangle + \widetilde{\tau}_{i+1}^{-1}\langle \partial F^*(y^{i+1}), y^{i+1} - \widehat{y}\rangle \\
&\qquad + \widetilde{\tau}_i^{-1}\langle K^* y^{i+1}, x^{i+1} - \widehat{x}\rangle - \widetilde{\tau}_{i+1}^{-1}\langle Kx^{i+1}, y^{i+1} - \widehat{y}\rangle
\end{aligned}$$

Since then $\Gamma_i \in ([0,\Gamma] + \widetilde{\mathcal{K}})/2$, and $R_{i+1} \in \hat{\mathcal{K}}/2$, we may take $\Gamma' = 2\Gamma_i$ and $R = 2R_{i+1}$ in (G-pc) and (F*-pc). It follows

$$\begin{aligned}
&\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \\
&\quad \ge \widetilde{\tau}_i^{-1}\big(G(x^{i+1}) - G(\widehat{x}) \\
&\qquad + \frac{1}{2}\|x^{i+1} - \widehat{x}\|^2_{2\Gamma_i} - \frac{1}{2}\psi_{2\Gamma_i-\Gamma}(x^{i+1} - \widehat{x})\big) \\
&\qquad + \widetilde{\tau}_{i+1}^{-1}\big(F^*(y^{i+1}) - F^*(\widehat{y}) + \frac{1}{2}\|y^{i+1} - \widehat{y}\|^2_{2R_{i+1}} \\
&\qquad - \frac{1}{2}\phi_{2R_{i+1}}(y^{i+1} - \widehat{y})\big) - \widetilde{\tau}_i^{-1}\langle y^{i+1}, K\widehat{x}\rangle \\
&\qquad + \widetilde{\tau}_{i+1}^{-1}\langle \widehat{y}, Kx^{i+1}\rangle + (\widetilde{\tau}_i^{-1} - \widetilde{\tau}_{i+1}^{-1})\langle y^{i+1}, Kx^{i+1}\rangle.
\end{aligned}$$

Using (2) and (24), we can make all of the factors '2' and '1/2' in this expression annihilate each other. With D_{i+1} as in (27) and $\lambda = 1/2$, we therefore have

$$\begin{aligned}
&\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \\
&\quad \ge \widetilde{\tau}_i^{-1}\Big(G(x^{i+1}) - G(\widehat{x}) + \langle \widehat{y}, Kx^{i+1}\rangle\Big) \\
&\qquad + \|x^{i+1} - \widehat{x}\|^2_{\widetilde{\tau}_i^{-1}\Gamma_i} \\
&\qquad + \widetilde{\tau}_{i+1}^{-1}\Big(F^*(y^{i+1}) - F^*(\widehat{y}) - \langle y^{i+1}, K\widehat{x}\rangle\Big) \\
&\qquad + \|y^{i+1} - \widehat{y}\|^2_{\widetilde{\tau}_{i+1}^{-1}R_{i+1}}
\end{aligned}$$

$$+(\widetilde{\tau}_i^{-1}-\widetilde{\tau}_{i+1}^{-1})\left(\langle y^{i+1}-\widehat{y},K(x^{i+1}-\widehat{x})\rangle -\langle\widehat{y},K\widehat{x}\rangle\right)-D_{i+1}.$$

A little bit of reorganisation and referral to (13) for the expansion of $S_i\bar{\Gamma}_i$ thus gives

$$\begin{aligned}&\langle H(u^{i+1}),S_i^*(u^{i+1}-\widehat{u})\rangle\\ &\geq\widetilde{\tau}_i^{-1}\left(G(x^{i+1})-G(\widehat{x})+\langle\widehat{y},Kx^{i+1}\rangle\right)\\ &\quad+\widetilde{\tau}_{i+1}^{-1}\left(F^*(y^{i+1})-F^*(\widehat{y})-\langle y^{i+1},K\widehat{x}\rangle\right)\\ &\quad-(\widetilde{\tau}_i^{-1}-\widetilde{\tau}_{i+1}^{-1})\langle\widehat{y},K\widehat{x}\rangle+\frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{S_i\bar{\Gamma}_i}-D_{i+1}.\end{aligned}\tag{29}$$

Let us write

$$\begin{aligned}\mathcal{G}_+^i(u^{i+1},\widehat{u}):=&\left(\widetilde{\tau}_i^{-1}G(x^{i+1})+\widetilde{\tau}_i^{-1}\langle\widehat{y},Kx^{i+1}\rangle-\widetilde{\tau}_i^{-1}F^*(\widehat{y})\right)\\ &-\left(\widetilde{\tau}_{i+1}^{-1}G(\widehat{x})+\widetilde{\tau}_{i+1}^{-1}\langle y^{i+1},K\widehat{x}\rangle-\widetilde{\tau}_{i+1}^{-1}F^*(y^{i+1})\right).\end{aligned}$$

Observing here the switches between the indices $i+1$ and i of the step length parameters in comparison with the last step of (29), we thus obtain

$$\begin{aligned}\langle H(u^{i+1}),S_i(u^{i+1}-\widehat{u})\rangle\geq\mathcal{G}_+^i(u^{i+1},\widehat{u})-\mathcal{G}_+^i(\widehat{u},\widehat{u})\\ +\frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{S_i\bar{\Gamma}_i}-D_{i+1}.\end{aligned}\tag{30}$$

We note that S_iM_i in (12) is self-adjoint as we have assumed T_i and Σ_{i+1} to be, and taken $\widetilde{T}_i$ and $\hat{T}_{i+1}$ to be scalars times the identity. We therefore deduce from the proof of Theorem 2.1 that (21) holds. Using (PP) to combine (21) and (30), we thus deduce

$$\begin{aligned}&\frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{S_{i+1}M_{i+1}}+\mathcal{G}_+^i(u^{i+1},\widehat{u})\\ &-\mathcal{G}_+^i(\widehat{u},\widehat{u})\leq\frac{1}{2}\|u^i-\widehat{u}\|^2_{S_iM_i}+D_{i+1}.\end{aligned}$$

Summing this for $i=0,\dots,N-1$ gives with C_0 from (27) the estimate

$$\begin{aligned}&\frac{1}{2}\|u^N-\widehat{u}\|^2_{S_NM_N}+\sum_{i=0}^{N-1}\left(\mathcal{G}_+^i(u^{i+1},\widehat{u})-\mathcal{G}_+^i(\widehat{u},\widehat{u})\right)\\ &\leq C_0+\sum_{i=0}^{N-1}D_{i+1}.\end{aligned}\tag{31}$$

We want to estimate the sum of the gaps $\mathcal{G}_+^i$ in (31). Using the convexity of G and F^*, we observe

$$\begin{aligned}&\sum_{i=0}^{N-1}\widetilde{\tau}_i^{-1}G(x^{i+1})\geq\widetilde{q}_NG(x_N),\quad\text{and}\\ &\sum_{i=0}^{N-1}\widetilde{\tau}_{i+1}^{-1}F^*(y^{i+1})\geq\hat{q}_NF^*(y_N).\end{aligned}\tag{32}$$

Also, by (25) and simple reorganisation

$$\sum_{i=0}^{N-1}\widetilde{\tau}_{i+1}^{-1}G(\widehat{x})=\widetilde{q}_NG(\widehat{x})+\widetilde{\tau}_N^{-1}G(\widehat{x})-\widetilde{\tau}_0^{-1}G(\widehat{x}),\quad\text{and}\tag{33}$$

$$\sum_{i=0}^{N-1}\widetilde{\tau}_i^{-1}F^*(\widehat{y})=\hat{q}_NF^*(y_N)-\widetilde{\tau}_N^{-1}F^*(\widehat{y})+\widetilde{\tau}_0^{-1}F^*(\widehat{y}).\tag{34}$$

All of (32)–(34) together give

$$\begin{aligned}&\sum_{i=0}^{N-1}\mathcal{G}_+^i(u^{i+1},\widehat{u})\\ &\geq\left(\widetilde{q}_NG(x_N)+\widetilde{q}_N\langle\widehat{y},Kx_N\rangle-\hat{q}_NF^*(\widehat{y})\right)\\ &\quad-\left(\widetilde{q}_NG(\widehat{x})+\hat{q}_N\langle y_N,K\widehat{x}\rangle-\hat{q}_NF^*(y_N)\right)\\ &\quad+\left(\widetilde{\tau}_N^{-1}G(\widehat{x})-\widetilde{\tau}_0^{-1}G(\widehat{x})+\widetilde{\tau}_N^{-1}F^*_{\hat{T}_N^{-1},*}(\widehat{x})-\widetilde{\tau}_0^{-1}F^*(\widehat{y})\right).\end{aligned}$$

Another use of (25) gives

$$\begin{aligned}&\sum_{i=0}^{N-1}\mathcal{G}_+^i(\widehat{u},\widehat{u})\\ &=(\widetilde{q}_N-\hat{q}_N)\langle\widehat{y},K\widehat{x}\rangle\\ &\quad+\left(\widetilde{\tau}_N^{-1}G(\widehat{x})-\widetilde{\tau}_0^{-1}G(\widehat{x})+\widetilde{\tau}_N^{-1}F^*(\widehat{x})-\widetilde{\tau}_0^{-1}F^*(\widehat{y})\right).\end{aligned}$$

Thus,

$$\sum_{i=0}^{N-1}\left(\mathcal{G}_+^i(u^{i+1},\widehat{u})-\mathcal{G}_+^i(\widehat{u},\widehat{u})\right)\geq\widetilde{q}_N\mathcal{G}^N+r_N,\tag{35}$$

where the remainder

$$r_N=(\widetilde{q}_N-\hat{q}_N)\left(F^*(\hat{y})-F^*(y_N)-\langle\hat{y}-y_N,K\hat{x}\rangle\right).$$

At a solution $\widehat{u}=(\widehat{x},\widehat{y})$ to (OC), we have $K\hat{x}\in\partial F^*(\hat{y})$, so $r_N\geq0$ provided $\widetilde{q}_N\leq\hat{q}_N$. But $\widetilde{q}_N-\hat{q}_N=\widetilde{\tau}_0^{-1}-\widetilde{\tau}_N^{-1}$, so this is guaranteed by our assumption (C3″). Using (35) in (31) therefore gives

$$\frac{1}{2}\|u^N-\widehat{u}\|^2_{S_NM_N}+\widetilde{q}_N\mathcal{G}^N+r_N\leq C_0+\sum_{i=0}^{N-1}D_{i+1}.\tag{36}$$

A referral to (C2) to estimate $S_N M_N$ from below shows (26), concluding the proof. □

4 Convergence Rates in Special Cases

To derive a practical algorithm, we need to satisfy the update rules (C1) and (C2), as well as the partial monotonicity conditions (G-PM) and (F*-PM). As we have already discussed in Sect. 3, this can be done when for some $\widetilde{\tau}_i > 0$ we set

$$\widetilde{T}_i = \widetilde{\tau}_i I, \quad \text{and} \quad \hat{T}_i = \widetilde{\tau}_i I. \tag{37}$$

The result of these choices is Algorithm 2, whose convergence we studied in Theorem 3.1. Our task now is to verify its conditions, in particular (G-pc) and (F*-pc) [alternatively (F*-pm) and (G-pm)], as well as (C1″), (C2″), and (C3″) for Γ of the projection form γP.

4.1 An Approach to Updating Σ

We have not yet defined an explicit update rule for Σ_{i+1}, merely requiring that it has to satisfy (C2″) and (C1″). The former in particular requires

$$\Sigma_{i+1}^{-1} \geq \widetilde{\omega}_i (1-\delta)^{-1} K T_i K^*.$$

Hiring the help of some linear operator $\mathcal{F} \in \mathcal{L}(\mathcal{L}(Y;Y); \mathcal{L}(Y;Y))$ satisfying

$$\mathcal{F}(K T_i K^*) \geq K T_i K^*, \tag{38}$$

our approach is to define

$$\Sigma_{i+1}^{-1} := \widetilde{\omega}_i (1-\delta)^{-1} \mathcal{F}(K T_i K^*). \tag{39}$$

Then, (C2″) is satisfied provided $T_i^{-1} \in \mathcal{Q}$. Since $\widetilde{\tau}_{i+1}^{-1} \Sigma_{i+1}^{-1} = \widetilde{\tau}_i^{-1} (1-\delta)^{-1} \mathcal{F}(K T_i K^*)$, the condition (C1″) reduces into the satisfaction for each $i \in \mathbb{N}$ of

$$\widetilde{\tau}_i^{-1}(I + 2\Gamma T_i) T_i^{-1} - \widetilde{\tau}_{i+1}^{-1} T_{i+1}^{-1} \geq -2\widetilde{\tau}_i^{-1}(\Gamma_i - \Gamma), \quad \text{and} \tag{40a}$$

$$\frac{1}{1-\delta}\left(\widetilde{\tau}_i^{-1} \mathcal{F}\left(K T_i K^*\right) - \widetilde{\tau}_{i+1}^{-1} \mathcal{F}\left(K T_{i+1} K^*\right)\right) \geq -2\widetilde{\tau}_{i+1}^{-1} R_{i+1}. \tag{40b}$$

To apply Theorem 3.1, all that remains is to verify in special cases the conditions (40) together with (C3″) and the partial strong convexity conditions (G-pc) and (F*-pc).

4.2 When Γ is a Multiple of a Projection

We now take $\Gamma = \bar{\gamma} P$ for some $\bar{\gamma} > 0$, and a projection operator $P \in \mathcal{L}(X;X)$: idempotent, $P^2 = P$, and self-adjoint, $P^* = P$. We let $P^\perp := I - P$. Then, $P^\perp P = P P^\perp = 0$. With this, we assume that $\widetilde{\mathcal{K}}$ is such that for some $\bar{\gamma}^\perp > 0$ holds

$$[0, \bar{\gamma}^\perp P^\perp] \subset \widetilde{\mathcal{K}}. \tag{41}$$

To unify our analysis for gap and non-gap estimates of Theorem 3.1, we now pick $\lambda = 1/2$ in the former case, and $\lambda = 1$ in the latter. We then pick $0 \leq \gamma \leq \lambda \bar{\gamma}$, and $0 \leq \gamma_i^\perp \leq \lambda \bar{\gamma}^\perp$, and set

$$T_i = \tau_i P + \tau_i^\perp P^\perp, \quad \Omega_i = \omega_i P + \omega_i^\perp P^\perp, \quad \text{and}$$
$$\Gamma_i = \gamma P + \gamma_i^\perp P^\perp. \tag{42}$$

With this, $\tau_i, \tau_i^\perp > 0$ guarantee $T_i \in \mathcal{Q}$. Moreover, T_i is self-adjoint. Moreover, $\Gamma_i \in \lambda([0,\Gamma] + \widetilde{\mathcal{K}})$, exactly as required in both the gap and the non-gap cases of Theorem 3.1.

Since

$$\begin{aligned} K T_i K^* &= \tau_i K P K^* + \tau_i^\perp K P^\perp K^* \\ &= (\tau_i - \tau_i^\perp) K P K^* + \tau_i^\perp K K^*, \end{aligned}$$

we are encouraged to take

$$\mathcal{F}(K T_i K^*) := \max\{0, \tau_i - \tau_i^\perp\} \|K P\|^2 I + \tau_i^\perp \|K\|^2 I. \tag{43}$$

Observe that (43) satisfies (38). Inserting (43) into (39), we obtain

$$\Sigma_{i+1} = \sigma_{i+1} I \quad \text{with}$$
$$\sigma_{i+1}^{-1} = \frac{\widetilde{\omega}_i}{1-\delta}\left(\max\{0, \tau_i - \tau_i^\perp\} \|K P\|^2 + \tau_i^\perp \|K\|^2\right). \tag{44}$$

Since Σ_{i+1} is now equivalent to a scalar, (40b), we also take $R_{i+1} = \rho_{i+1} I$, assuming for some $\bar{\rho} > 0$ that

$$[0, \bar{\rho} I] \subset \hat{\mathcal{K}}.$$

Setting

$$\eta_i := \widetilde{\tau}_i^{-1} \max\{0, \tau_i - \tau_i^\perp\} - \widetilde{\tau}_{i+1}^{-1} \max\{0, \tau_{i+1} - \tau_{i+1}^\perp\}$$

we thus expand (40) as

$$\widetilde{\tau}_i^{-1}(1 + 2\gamma\tau_i)\tau_i^{-1} - \widetilde{\tau}_{i+1}\tau_{i+1}^{-1} \geq 0, \tag{45a}$$
$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1} \geq -2\widetilde{\tau}_i^{-1}\gamma_i^\perp, \tag{45b}$$

$$\frac{1}{1-\delta}\left(\eta_i\|KP\|^2+(\widetilde{\tau}_i^{-1}\tau_i^{\perp}-\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp})\|K\|^2\right)\geq -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \tag{45c}$$

We are almost ready to state a general convergence result for projective Γ. However, we want to make one more thing more explicit. Since the choices (42) satisfy

$$\Gamma_i-\lambda\Gamma=(\gamma-\lambda\bar{\gamma})P+\gamma_i^{\perp}P^{\perp}\leq\gamma_i^{\perp}P^{\perp} \quad \text{and} \quad R_{i+1}=\rho_{i+1}I,$$

we suppose for simplicity that

$$\psi_{\Gamma_i-\lambda\Gamma}(x)=\gamma_i^{\perp}\psi^{\perp}(P^{\perp}x) \quad \text{and} \quad \phi_{R_{i+1}}(y)=\rho_{i+1}\phi(y) \tag{46}$$

for some $\psi^{\perp}:P^{\perp}X\to\mathbb{R}$ and $\phi:Y\to\mathbb{R}$. The conditions (G-pc) and (F*-pc) reduce in this case to the satisfaction for some $\bar{\gamma},\bar{\gamma}^{\perp},\bar{\rho}>0$ of

$$G(x')-G(x)\geq\langle\partial G(x),x'-x\rangle+\frac{\bar{\gamma}}{2}\|P(x'-x)\|^2+\frac{\gamma^{\perp}}{2}\left(\|P^{\perp}(x'-x)\|^2-\psi(P^{\perp}(x'-x))\right), \tag{G-pcr}$$

for all $x,x'\in X$ and $0\leq\gamma^{\perp}\leq\bar{\gamma}^{\perp}$, as well as of

$$F^*(y')-F^*(y)\geq\langle\partial F^*(y),y'-y\rangle+\frac{\rho}{2}\left(\|y'-y\|^2-\phi(y'-y)\right), \tag{F*-pcr}$$

for all $y,y'\in Y$ and $0\leq\rho\leq\bar{\rho}$. Analogues of (G-pm) and (F*-pm) can be formed.

To summarise the findings of this section, we state the following proposition.

Proposition 4.1 *Suppose* (G-pcr) *and* (F*-pcr) *hold for some projection operator* $P\in\mathcal{L}(X;X)$ *and scalars* $\bar{\gamma},\bar{\gamma}^{\perp},\bar{\rho}>0$. *With* $\lambda=1/2$, *pick* $\gamma\in[0,\lambda\bar{\gamma}]$. *For each* $i\in\mathbb{N}$, *suppose* (45) *is satisfied with*

$$0\leq\gamma_i^{\perp}\leq\lambda\bar{\gamma}^{\perp},\quad 0\leq\rho_i\leq\lambda\bar{\rho},\quad \text{and}\quad \widetilde{\tau}_0\geq\widetilde{\tau}_i>0. \tag{47}$$

If we solve (45a) *exactly, define* T_i, Γ_i, *and* Σ_{i+1} *through* (42) *and* (44), *and set* $R_{i+1}=\rho_{i+1}I$, *then the iterates of Algorithm 2 satisfy with* C_0 *and* D_{i+1} *as in* (27) *the estimate*

$$\frac{\delta}{2}\|P(x^N-\widehat{x})\|^2+\frac{1}{\tau_0^{-1}+2\gamma}\mathcal{G}^N\leq\widetilde{\tau}_N\tau_N\left(C_0+\sum_{i=0}^{N-1}D_{i+1}\right). \tag{48}$$

If we take $\lambda=1$, *then* (48) *holds with* $\mathcal{G}^N=0$.

Observe that presently

$$D_{i+1}=\widetilde{\tau}_i^{-1}\gamma_i^{\perp}\psi^{\perp}(P^{\perp}(x^{i+1}-\widehat{x}))+\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}\phi(y^{i+1}-\widehat{y}). \tag{49}$$

Proof As we have assumed through (47), or otherwise already verified its conditions, we may apply Theorem 3.1. Multiplying (26) by $\widetilde{\tau}_N\tau_N$, we obtain

$$\frac{\delta}{2}\|x^N-\widehat{x}\|_P^2+\widetilde{q}_N\widetilde{\tau}_N\tau_N\mathcal{G}^N\leq\widetilde{\tau}_N\tau_N\left(C_0+\sum_{i=0}^{N-1}D_{i+1}\right). \tag{50}$$

Now, observe that solving (45a) exactly gives

$$\widetilde{\tau}_N^{-1}\tau_N^{-1}=\widetilde{\tau}_{N-1}^{-1}\tau_{N-1}^{-1}+2\gamma\widetilde{\tau}_{N-1}^{-1}=\widetilde{\tau}_0^{-1}\tau_0^{-1}+\sum_{j=0}^{N-1}2\gamma\widetilde{\tau}_j^{-1}=\widetilde{\tau}_0^{-1}\tau_0^{-1}+2\gamma\widetilde{q}_N. \tag{51}$$

Therefore, we have the estimate

$$\widetilde{q}_N\widetilde{\tau}_N\tau_N=\frac{\widetilde{q}_N}{\widetilde{\tau}_0^{-1}\tau_0^{-1}+2\gamma\widetilde{q}_N}=\frac{1}{\widetilde{\tau}_0^{-1}\tau_0^{-1}\widetilde{q}_N^{-1}+2\gamma}\geq\frac{1}{\tau_0^{-1}+2\gamma}. \tag{52}$$

With this, (50) yields (48). □

4.3 Primal and Dual Penalties with Projective Γ

We now study conditions that guarantee the convergence of the sum $\widetilde{\tau}_N\tau_N\sum_{i=0}^{N-1}D_{i+1}$ in (48). Indeed, the right-hand sides of (45b) and (45c) relate to D_{i+1}. In most practical cases, which we study below, ϕ and ψ transfer these right-hand side penalties into simple linear factors within D_{i+1}. Optimal rates are therefore obtained by solving (45b) and (45c) as equalities, with the right-hand sides proportional to each other. Since $\eta_i\geq 0$, and it will be the case that $\eta_i=0$ for large i, we, however, replace (45c) by the simpler condition

$$\frac{1}{1-\delta}(\widetilde{\tau}_i^{-1}\tau_i^{\perp}-\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp})\|K\|^2\geq -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \tag{53}$$

Then, we try to make the left-hand sides of (45b) and (53) proportional with only $\tau_{i+1}^{\perp}$ as a free variable. That is, for some proportionality constant $\zeta>0$, we solve

$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1}-\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1}=\zeta(\widetilde{\tau}_i^{-1}\tau_i^{\perp}-\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp}). \tag{54}$$

Multiplying both sides of (54) by $\zeta^{-1}\widetilde{\tau}_{i+1}\tau_{i+1}^{\perp}$, gives on $\tau_{i+1}^{\perp}$ the quadratic condition

$$\tau_{i+1}^{\perp,2} + \widetilde{\omega}_i(\zeta^{-1}\tau_i^{\perp,-1} - \tau_i^{\perp})\tau_{i+1}^{\perp} - \zeta^{-1} = 0.$$

Thus,

$$\tau_{i+1}^{\perp} = \frac{1}{2}\left(\widetilde{\omega}_i(\tau_i^{\perp} - \zeta^{-1}\tau_i^{\perp,-1}) + \sqrt{\widetilde{\omega}_i^2(\tau_i^{\perp} - \zeta^{-1}\tau_i^{\perp,-1})^2 + 4\zeta^{-1}}\right). \tag{55}$$

Solving (45b) and (53) as equalities, (54) and (55) give

$$2\widetilde{\tau}_i^{-1}\gamma_i^{\perp} = \frac{2\zeta(1-\delta)}{\|K\|^2}\widetilde{\tau}_{i+1}^{-1}\rho_{i+1} = \zeta(\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp} - \widetilde{\tau}_i^{-1}\tau_i^{\perp}). \tag{56}$$

Note that this quantity is non-negative exactly when $\omega_i^{\perp} \geq \widetilde{\omega}_i$. We have

$$\begin{aligned}\frac{\omega_i^{\perp}}{\widetilde{\omega}_i} &= \frac{\tau_{i+1}^{\perp}}{\tau_i^{\perp}\widetilde{\omega}_i} \\ &= \frac{1}{2}\left(1 - \zeta^{-1}\tau_i^{\perp,-2} + \sqrt{(1-\zeta^{-1}\tau_i^{\perp,-2})^2 + 4\zeta^{-1}\widetilde{\omega}_i^{-2}\tau_i^{\perp,-2}}\right).\end{aligned}$$

This quickly yields $\omega_i^{\perp} \geq \widetilde{\omega}_i$ if $\widetilde{\omega}_i \leq 1$. In particular, (56) is non-negative when $\widetilde{\omega}_i \leq 1$.

The next lemma summarises these results for the standard choice of $\widetilde{\omega}_i$.

Lemma 4.1 *Let $\tau_{i+1}^{\perp}$ by given by (55), and set*

$$\widetilde{\omega}_i = \omega_i = 1/\sqrt{1+2\gamma\tau_i}. \tag{57}$$

Then, $\omega_i^{\perp} \geq \widetilde{\omega}_i$, $\widetilde{\tau}_i \leq \widetilde{\tau}_0$, and (45) is satisfied with the right-hand sides given by the non-negative quantity in (56). Moreover,

$$\tau_i^{\perp} \leq \zeta^{-1/2} \implies \tau_{i+1}^{\perp} \leq \zeta^{-1/2}. \tag{58}$$

Proof The choice (57) satisfies (45a), so that (45) in its entirety will be satisfied with the right-hand sides of (45b)–(45c) given by (56). The bound $\widetilde{\tau}_i \leq \widetilde{\tau}_0$ follows from $\widetilde{\omega}_i \leq 1$. Finally, the implication (58) is a simple estimation of (55). □

Specialisation of Algorithm 2 to the choices in Lemma 4.1 yields the steps of Algorithm 3. Observe that $\widetilde{\tau}_i$ entirely disappears from the algorithm. To obtain convergence rates, and to justify the initial conditions, we will shortly seek to exploit with specific ϕ and ψ the telescoping property stemming from the non-negativity of the last term of (56).

Algorithm 3 Partial acceleration for projective Γ—primal and dual penalties

Require: F^* and G satisfying (G-pcr) and (F*-pcr) for some $\bar{\gamma}, \bar{\gamma}^{\perp}, \bar{\rho} \geq 0$, and a projection operator $P \in \mathcal{L}(X;X)$. A choice of $\gamma \in [0, \bar{\gamma}]$. Initial step length parameters $\tau_0, \tau_0^{\perp} > 0$, a choice of $\delta \in (0,1)$, and $\zeta \leq \tau_0^{\perp,-2}$, all satisfying (61).

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **repeat**
3: Set
$\omega_i = 1/\sqrt{1+2\gamma\tau_i}$, and
$\omega_i^{\perp} = \frac{1}{2}\left((1-\zeta^{-1}\tau_i^{\perp,-2})\omega_i + \sqrt{(1-\zeta^{-1}\tau_i^{\perp,-2})^2\omega_i^2 + 4\zeta^{-1}\tau_i^{\perp,-2}}\right)$.
4: Update
$\tau_{i+1} = \tau_i\omega_i$, $\tau_{i+1}^{\perp} = \tau_i^{\perp}\omega_i^{\perp}$, and
$\sigma_{i+1} = \omega_i^{-1}(1-\delta)/(\max\{0, \tau_i - \tau_i^{\perp}\}\|KP\|^2 + \tau_i^{\perp}\|K\|^2)$,
5: With $T_i = \tau_i P + \tau_i^{\perp}P^{\perp}$, perform the updates

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_iK^*y^i), \tag{59a}$$
$$\bar{x}^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \tag{59b}$$
$$y^{i+1} := (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1}K\bar{x}^{i+1}). \tag{59c}$$

6: **until** a stopping criterion is fulfilled.

There is still, however, one matter to take care of. We need $\rho_i \leq \lambda\bar{\rho}$ and $\gamma_i^{\perp} \leq \lambda\bar{\gamma}^{\perp}$, although in many cases of practical interest, the upper bounds are infinite and hence inconsequential. We calculate from (55) and (57) that

$$\begin{aligned}\gamma_i^{\perp} &= \frac{\zeta}{2}(\widetilde{\omega}_i^{-1}\tau_{i+1}^{\perp} - \tau_i^{\perp}) = \frac{1}{2}\left(-\zeta\tau_i^{\perp} - \tau_i^{\perp,-1} + \sqrt{(\zeta\tau_i^{\perp} - \tau_i^{\perp,-1})^2 + 4\zeta\widetilde{\omega}_i^{-2}}\right) \\ &\leq \sqrt{\zeta(\widetilde{\omega}_i^{-2} - 1)} = \sqrt{2\zeta\gamma\tau_i} \leq \sqrt{2\zeta\gamma\tau_0}.\end{aligned} \tag{60}$$

Therefore, we need to choose ζ and τ_0 to satisfy $2\zeta\gamma\tau_0 \leq (\lambda\bar{\gamma}^{\perp})^2$. Likewise, we calculate from (56), (57), and (60) that

$$\begin{aligned}\rho_{i+1} &= \frac{\widetilde{\omega}_i}{c}\gamma_i^{\perp} = \frac{\|K\|^2\widetilde{\omega}_i}{(1-\delta)\zeta}\gamma_i^{\perp} \leq \frac{\|K\|^2\widetilde{\omega}_i}{(1-\delta)\zeta}\sqrt{2\zeta\gamma\tau_i} \\ &= \frac{\|K\|^2}{(1-\delta)\zeta}\sqrt{2\zeta\gamma\tau_0}.\end{aligned}$$

This tells us to choose τ_0 and ζ to satisfy $2\|K\|^4/(1-\delta)^2\zeta^{-1}\gamma\tau_0 \leq (\lambda\bar{\rho})^2$. Overall, we obtain for τ_0 and ζ the condition

$$0 < \tau_0 \leq \frac{\lambda^2}{2\gamma}\min\left\{\frac{\bar{\gamma}^{\perp,2}}{\zeta}, \frac{\bar{\rho}^2\zeta(1-\delta)^2}{\|K\|^4}\right\}. \tag{61}$$

This can always be satisfied through suitable choices of τ_0 and ζ.

If now $\phi \equiv C_\phi$ and $\psi \equiv C_\psi^\perp$, using the non-negativity of (56), we calculate

$$\sum_{i=0}^{N-1} \widetilde{\tau}_{i+1}^{-1}\rho_{i+1}\phi(y^{i+1}-\widehat{y}) = \frac{\|K\|^2 C_\phi}{2(1-\delta)} \sum_{i=0}^{N-1} \left(\frac{\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^\perp}{2} - \frac{\widetilde{\tau}_i^{-1}\tau_i^\perp}{2}\right) \leq \frac{\|K\|^2 C_\phi}{2(1-\delta)}\widetilde{\tau}_N^{-1}\tau_N^\perp. \tag{62}$$

Similarly

$$\sum_{i=0}^{N-1} \widetilde{\tau}_i^{-1}\gamma_i^\perp \psi(x^{i+1}-\widehat{x}) \leq \frac{\zeta C_\psi^\perp}{2}\widetilde{\tau}_N^{-1}\tau_N^\perp. \tag{63}$$

Using these expression to expand (49), we obtain the following convergence result.

Theorem 4.1 *Suppose* (G-pcr) *and* (F*-pcr) *hold for some projection operator* $P \in \mathcal{L}(X;X)$, *scalars* $\bar{\gamma}, \bar{\gamma}^\perp, \bar{\rho} > 0$ *with* $\phi \equiv C_\phi$, *and* $\psi \equiv C_\psi^\perp$, *for some constants* $C_\phi, C_\psi^\perp > 0$. *With* $\lambda = 1/2$, *fix* $\gamma \in (0, \lambda\bar{\gamma}]$. *Select initial* $\tau_0, \tau_0^\perp > 0$, *as well as* $\delta \in (0,1)$ *and* $\zeta \leq (\tau_0^\perp)^{-2}$ *satisfying* (61). *Then, Algorithm 3 satisfies for some* $C_0, C_\tau > 0$ *the estimate*

$$\frac{\delta}{2}\|P(x^N-\widehat{x})\|^2 + \frac{1}{\tau_0^{-1}+2\gamma}\mathcal{G}^N \leq \frac{C_0 C_\tau^2}{N^2} + \frac{C_\tau}{2N}\left(\zeta^{1/2}C_\psi^\perp + \frac{\zeta^{-1/2}\|K\|^2}{1-\delta}C_\phi\right), \quad (N \geq 0). \tag{64}$$

If we take $\lambda = 1$, *then* (48) *holds with* $\mathcal{G}^N = 0$.

Proof During the course of the derivation of Algorithm 3, we have verified (45), solving (45a) as an equality. Moreover, Lemma 4.1 and (61) guarantee (47). We may therefore apply Proposition 4.1. Inserting (62) and (63) into (48) and (49) gives

$$\frac{\delta}{2}\|P(x^N-\widehat{x})\|^2 + \frac{1}{\tau_0^{-1}+2\gamma}\mathcal{G}^N \leq \tau_N\widetilde{\tau}_N \times \left(C_0 + \frac{\zeta C_\psi^\perp}{2}\widetilde{\tau}_N^{-1}\tau_N^\perp + \frac{\|K\|^2 C_\phi}{2(1-\delta)}\widetilde{\tau}_N^{-1}\tau_N^\perp\right). \tag{65}$$

The condition $\zeta \leq (\tau_0^\perp)^{-2}$ now guarantees $\tau_N^\perp \leq \zeta^{-1/2}$ through (58). Now we note that $\widetilde{\tau}_i$ is not used in Algorithm 3, so it only affects the convergence rate estimates. We therefore simply take $\widetilde{\tau}_0 = \tau_0$, so that $\widetilde{\tau}_N = \tau_N$ for all $N \in \mathbb{N}$. With this and the bound $\tau_N \leq C_\tau/N$ from Remark 3.2, (64) follows by simple estimation of (65). □

Remark 4.1 As a special case of Algorithm 3, if we choose $\zeta = \tau_0^{\perp,-2}$, then we can show from (55) that $\tau_i^\perp = \tau_0^\perp = \zeta^{-1/2}$ for all $i \in \mathbb{N}$.

Remark 4.2 The convergence rate provided by Theorem 4.1 is a mixed $O(1/N^2) + O(1/N)$ rate, similarly to that derived in [5] for a type of forward–backward splitting algorithm for smooth G. Ours is of course backward–backward type algorithm. It is interesting to note that using the differentiability properties of infimal convolutions [23, Proposition 18.7], and the presentation of a smooth G as an infimal convolution, it is formally possible to derive a forward–backward algorithm from Algorithm 3. The difficulties lie in combining this conversion trick with conditions on the step lengths.

4.4 Dual Penalty Only with Projective Γ

Continuing with the projective Γ setup of Sect. 4.2, we now study the case $\widetilde{\mathcal{K}} = \{0\}$, that is, when only the dual penalty ϕ is available with $\psi \equiv 0$. To use Proposition 4.1, we need to satisfy (47) and (45), with (45a) holding as an equality. Since $\gamma_i^\perp = 0$, (45b) becomes

$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1} \geq 0. \tag{66}$$

With respect to $\tau_{i+1}^\perp$, the left-hand side of (45c) is maximised (and the penalty on the right-hand side minimised) when (66) is minimised. Thus, we solve (66) exactly, which gives

$$\tau_{i+1}^\perp = \tau_i^\perp \widetilde{\omega}_i^{-1}.$$

In consequence $\omega_i^\perp = \widetilde{\omega}_i^{-1}$, and (45c) becomes

$$\frac{1}{1-\delta}\eta_i\|KP\|^2 + \frac{\widetilde{\tau}_i^{-2}}{1-\delta}(1-\widetilde{\omega}_i^{-2})\|K\|^2 \geq -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \tag{67}$$

In order to simultaneously satisfy (45a), this suggests for some, yet undetermined, $a_i > 0$, to choose

$$\widetilde{\omega}_i := \frac{1}{\sqrt{1+a_i\widetilde{\tau}_i^2}} \quad \text{and} \quad \omega_i := \frac{1}{\widetilde{\omega}_i(1+2\gamma\tau_i)}. \tag{68}$$

Since $\eta_i \geq 0$, (67) is satisfied with the choice (68) if we take

$$\rho_{i+1} = \widetilde{\tau}_{i+1}a_i\frac{\|K\|^2}{2(1-\delta)}.$$

To use Proposition 4.1, we need to satisfy $\rho_{i+1} \leq \lambda\bar{\rho}$. Since (68) implies that $\{\widetilde{\tau}_i\}_{i=0}^\infty$ is non-increasing, we can satisfy this for large enough i if $a_i \searrow 0$. To ensure satisfaction for all $i \in \mathbb{N}$, it suffices to take $\{a_i\}_{i=0}^\infty$ non-increasing, and satisfy the initial condition

$$a_0\widetilde{\tau}_0\frac{\|K\|^2}{2(1-\delta)} \leq \lambda\bar{\rho}. \tag{69}$$

The rule $\widetilde{\tau}_{i+1} = \widetilde{\omega}_i \widetilde{\tau}_i$ and (68) give $\widetilde{\tau}_{i+1}^{-2} = \widetilde{\tau}_i^{-2} + a_i$. We therefore see that

$$\widetilde{\tau}_N^{-1}\tau_N^{-1} = \widetilde{\tau}_0^{-1}\tau_0^{-1} + 2\gamma \sum_{i=0}^{N-1} \sqrt{\widetilde{\tau}_0^{-2} + \textstyle\sum_{j=0}^{i-1} a_j} \geq 2\gamma \sum_{i=0}^{N-1} \sqrt{\widetilde{\tau}_0^{-2} + \textstyle\sum_{j=0}^{i-1} a_j} =: 1/\mu_0^N.$$

Assuming ϕ to have the structure (46), moreover,

$$\sum_{i=0}^{N-1} D_{i+1} = \sum_{i=0}^{N-1} \phi_{\widetilde{\tau}_{i+1}^{-1} R_{i+1}}(y^{i+1} - \widehat{y}) = \frac{\|K\|^2}{2(1-\delta)} \sum_{i=0}^{N-1} a_i \phi(y^{i+1} - \widehat{y}).$$

Thus, the rate (48) in Proposition 4.1 states

$$\frac{\delta}{2}\|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + 2\gamma}\mathcal{G}^N \leq \mu_0^N C_0 + \frac{\|K\|^2}{2(1-\delta)}\mu_1^N \tag{70}$$

for

$$\mu_1^N := \mu_0^N \sum_{i=0}^{N-1} a_i \phi(y^{i+1} - \widehat{y}).$$

The convergence rate is thus completely determined by μ_0^N and μ_1^N.

Remark 4.3 If $\phi \equiv 0$, that is, if F^* is strongly convex, we may simply pick $\widetilde{\omega}_i = \omega_i = 1/\sqrt{1 + 2\gamma\tau_i}$, that is $a_i = 2\gamma$, and obtain from (70) a $O(1/N^2)$ convergence rate.

For a more generally applicable algorithm, suppose $\phi(y^{i+1} - \widehat{y}) \equiv C_\phi$ as in Theorem 4.1. We need to choose a_i. One possibility is to pick some $q \in (0, 1]$ and

$$a_i := \widetilde{\tau}_0^{-2}\big((i+1)^q - i^q\big). \tag{71}$$

The concavity of $i \mapsto q^i$ for $q \in (0, 1]$ easily shows that $\{a_i\}_{i=0}^\infty$ is non-increasing. With the choice (71), we then compute

$$\sum_{i=0}^{N-1} \sqrt{\widetilde{\tau}_0^{-2} + \textstyle\sum_{j=0}^{i-1} a_j} = \widetilde{\tau}_0^{-1} \sum_{i=0}^{N-1} i^{q/2} \geq \widetilde{\tau}_0^{-1} \int_0^{N-1} x^{q/2}\,dx = \frac{\widetilde{\tau}_0^{-1}}{1 + q/2}(N-1)^{1+q/2},$$

and

$$\sum_{i=0}^{N-1} a_i \leq \widetilde{\tau}_0^{-2} N^q.$$

If $N \geq 2$, we find with $C_a = (1 + q/2)/(2^{1+q/2}\lambda\gamma)$ that

$$\mu_0^N \leq \frac{\widetilde{\tau}_0 C_a}{N^{1+q/2}}, \quad \text{and} \quad \mu_1^N \leq \frac{C_a C_\phi}{\widetilde{\tau}_0 N^{1-q/2}}. \tag{72}$$

The choice $q = 0$ gives uniform $O(1/N)$ over both the initialisation and the dual sequence. By choosing $q = 1$, we get $O(1/N^{3/2})$ convergence with respect to the initialisation, and $O(1/N^{1/2})$ with respect to the residual sequence.

With these choices, Algorithm 2 yields Algorithm 4, whose convergence properties are stated in the next theorem.

Algorithm 4 Partial acceleration for projective Γ—dual penalty only

Require: G satisfying (G-pcr) (with $\psi \equiv 0$) for some $\bar{\gamma} > 0$ and a projection operator $P \in \mathcal{L}(X; X)$. F^* satisfying (F*-pcr) for some $\bar{\rho} > 0$. A choice of $\gamma \in [0, \bar{\gamma}]$ and a non-increasing sequence $\{a_i\}_{i=0}^\infty$, for example as in (71). Initial step parameters $\tau_0, \tau_0^\perp, \widetilde{\tau}_0 > 0$, as well as $\delta \in (0, 1)$, satisfying (69).

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **repeat**
3: Set

$$\widetilde{\omega}_i := 1/\sqrt{1 + a_i\widetilde{\tau}_i^2}, \quad \widetilde{\tau}_{i+1} := \widetilde{\tau}_i\widetilde{\omega}_i, \quad \tau_{i+1}^\perp := \tau_i^\perp/\widetilde{\omega}_i,$$
$$\omega_i := \widetilde{\omega}_i^{-1}/(1 + 2\gamma\tau_i), \quad \tau_{i+1} := \tau_i\omega_i,$$

as well as

$$\sigma_{i+1} = \omega_i^{-1}(1-\delta)/(\max\{0, \tau_i - \tau_i^\perp\}\|KP\|^2 + \tau_i^\perp\|K\|^2).$$

4: With $T_i := \tau_i P + \tau_i^\perp P^\perp$, perform the updates

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_i K^* y^i),$$
$$\bar{x}^{i+1} := \widetilde{\omega}_i(x^{i+1} - x^i) + x^{i+1},$$
$$y^{i+1} := (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1} K\bar{x}^{i+1}).$$

5: **until** a stopping criterion is fulfilled.

Theorem 4.2 *Suppose* (G-pcr) *and* (F*-pcr) *hold for some projection operator* $P \in \mathcal{L}(X; X)$ *and* $\bar{\gamma}, \bar{\gamma}^\perp, \bar{\rho} \geq 0$ *with* $\psi \equiv 0$ *and* $\phi \equiv C_\phi$ *for some constant* $C_\phi \geq 0$. *With* $\lambda = 1/2$, *choose* $\gamma \in (0, \lambda\bar{\gamma}]$, *and pick the sequence* $\{a_i\}_{i=0}^\infty$ *by* (71) *for some* $q \in (0, 1]$. *Select initial* $\tau_0, \tau_0^\perp, \widetilde{\tau}_0 > 0$ *and* $\delta \in (0, 1)$ *verifying* (69). *Then, Algorithm 4 satisfies*

$$\frac{\delta}{2}\|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + \gamma}\mathcal{G}^N \leq \frac{\widetilde{\tau}_0 C_a C_0}{N^{1+q/2}} + \frac{C_a C_\phi \|K\|^2}{2(1-\delta)\widetilde{\tau}_0^2 N^{1-q/2}}, \quad (N \geq 2). \tag{74}$$

If we take $\lambda = 1$, *then* (74) *holds with* $\mathcal{G}^N = 0$.

Proof We apply Proposition 4.1 whose assumptions we have verified during the course of the present section. In particular, $\widetilde{\tau}_i \leq \widetilde{\tau}_0$ through the choice (68) that forces $\widetilde{\omega}_i \leq 1$. Also, have already derived the rate (70) from (48). Inserting (72) into (70), noting that the former is only valid for $N \geq 2$, immediately gives (74). □

5 Examples from Image Processing and the Data Sciences

We now consider several applications of our algorithms. We generally have to consider discretisations, since many interesting infinite-dimensional problems necessitate Banach spaces. Using Bregman distances, it would be possible to generalise our work form Hilbert spaces to Banach spaces, as was done in [24] for the original method of [3]. This is, however, outside the scope of the present work.

5.1 Regularised Least Squares

A large range of interesting application problems can be written in the *Tikhonov regularisation* or *empirical loss minimisation* form

$$\min_{x \in X} G_0(f - Ax) + \alpha F(Kx). \tag{75}$$

Here $\alpha > 0$ is a regularisation parameter, $G_0 : Z \to \mathbb{R}$ typically convex and smooth fidelity term with data $f \in Z$. The forward operator $A \in \mathcal{L}(X; Z)$—which can often also be data—maps our unknown to the space of data. The operator $K \in \mathcal{L}(X; Y)$ and the typically non-smooth and convex $F : Y \to \overline{\mathbb{R}}$ act as a regulariser.

We are particularly interested in strongly convex G_0 and A with a non-trivial null-space. Examples include, for example, Lasso—a type of regularised regression—with $G_0 = \|x\|_2^2/2$, $K = I$, and $F(x) = \|x\|_1$, on finite-dimensional spaces. If the data of the Lasso is 'sparse', in the sense that A has a non-trivial null-space, then, based on accelerating the strongly convex part of the variable, our algorithm can provide improved convergence rates compared to standard non-accelerated methods.

In image processing examples abound, we refer to [25] for an overview. In total variation (TV) regularisation, we still take $F(x) = \|x\|_1$, but is $K = \nabla$ the gradient operator. Strictly speaking, this has to be formulated in the Banach space $\mathrm{BV}(\Omega)$, but we will consider the discretised setting to avoid this problem. For denoising of Gaussian noise with TV regularisation, we take $A = I$, and again $G_0 = \|x\|_2^2/2$. This problem is not so interesting to us, as it is fully strongly convex. In a simple form of TV inpainting—filling in missing regions of an image—we take A as a subsampling operator S mapping an image $x \in L^2(\Omega)$ to one in $L^2(\Omega \setminus \Omega_d)$, for $\Omega_d \subset \Omega$ the defect region that we want to recreate. Observe that in this case, $\Gamma = S^*S$ is directly a projection operator. This is therefore a problem for our algorithms! Related problems include reconstruction from subsampled magnetic resonance imaging (MRI) data (see, for example, [11,26]), where we take $A = S\mathfrak{F}$ for $\mathfrak{F}$ the Fourier transform. Still, A^*A is a projection operator, so the problem perfectly suits our algorithms.

Another related problem is total variation deblurring, where A is a convolution kernel. This problem is slightly more complicated to handle, as A^*A is not a projection operator. Assuming periodic boundary conditions on a box $\Omega = \prod_{i=1}^m [c_i, d_i]$, we can write $A = \mathfrak{F}^*\hat{a}\mathfrak{F}$, multiplying the Fourier transform by some $\hat{a} \in L^2(\Omega)$. If $|\hat{a}| \geq \gamma$ on a subdomain, we obtain a projection form Γ (it would also be possible to extend our theory to non-constant γ, but we have decided not to extend the length of the paper by doing so. Dualisation likewise provides a further alternative).

Satisfaction of convexity conditions In all of the above examples, when written in the saddle point form (P), F^* is a simple pointwise ball constraint. Lemma 2.1 thus guarantees (F*-pcr). If $F(x) = \|x\|_1$ and $K = I$, then clearly $\|P^\perp \hat{x}\|$ can be bounded in $Z = L^1$ for $\hat{x}$ the optimal solution to (75). Thus, for some $M > 0$, we can add to (75) the artificial constraint

$$G'(x) := \iota_{\|\cdot\|_Z \leq M}(P^\perp x). \tag{76}$$

In finite dimensions, this gives a bound in L^2. Lemma 2.1 gives (G-pcr) with $\bar{\gamma}^\perp = \infty$.

In case of our total variation examples, $F(x) = \|x\|_1$ and $K = \nabla$. Provided mean-zero functions are not in the kernel of A, one can through Poincar's inequality [27] on $\mathrm{BV}(\Omega)$ and a two-dimensional connected domain $\Omega \subset \mathbb{R}^2$ show that even the original infinite-dimensional problems have bounded solutions in $L^2(\Omega)$. We may therefore again add the artificial constraint (76) with $Z = L^2$ to (75).

Dynamic bounds and pseudo-duality gaps We seldom know the exact bound M, but can derive conservative estimates. Nevertheless, adding such a bound to Algorithm 4 is a simple, easily implemented projection of $P^\perp(x^i - T_i K^* y^i)$ into the constraint set. In practise, we do not use or need the projection, and update the bound M dynamically so as to ensure that the constraint (76) is never active. Indeed, A having a non-trivial nullspace also causes duality gaps for (P) to be numerically infinite. In [28], a 'pseudo-duality gap' was therefore introduced, based on dynamically updating M. We will also use this type of dynamic duality gaps in our reporting.

Fig. 1 We use sample image (**b**) for denoising, and (**c**) for deblurring experiments. Free Kodak image suite photo, at the time of writing online at http://r0k.us/graphics/kodak/. **a** True image. **b** Noise image. **c** Blurry image

5.2 TGV^2 Regularised Problems

So far, we have considered very simple regularisation terms. Total generalised variation, TGV, was introduced in [29] as a higher-order generalisation of TV. It avoids the unfortunate stair-casing effect of TV—large flat areas with sharp transitions—while preserving the critical edge preservation property that smooth regularisers lack. We concentrate on the second-order TGV^2. In all of our image processing examples, we can replace TV by TGV^2.

As with total variation, we have to consider discretised models due the original problem being set in the Banach space $\mathrm{BV}(\Omega)$. For two parameters $\alpha, \beta > 0$, the regularisation functional is written in the differentiation cascade form of [30] as

$$\mathrm{TGV}^2_{(\beta,\alpha)}(u) := \min_w \ \alpha\|\nabla u - w\|_1 + \beta\|\mathcal{E}u\|_1.$$

Here $\mathcal{E} = (\nabla^T + \nabla)/2$ is the symmetrised gradient. With $x = (v, w)$ and $y = (y_1, y_2)$, we may write the problem

$$\min_v G_0(f - Av) + \mathrm{TGV}^2_{(\beta,\alpha)}(v), \tag{77}$$

in the saddle point form (P) with

$$\begin{aligned} G(x) &:= G_0(f - Av), \\ F^*(y) &= \iota_{\|\cdot\|_{L^\infty} \le \alpha}(y_1) + \iota_{\|\cdot\|_{L^\infty} \le \beta}(y_2), \quad \text{and} \\ K &:= \begin{pmatrix} \nabla & -I \\ 0 & \mathcal{E} \end{pmatrix}. \end{aligned}$$

If $A = I$, as is the case for denoising, we have

$$\Gamma = \gamma P \quad \text{for} \quad P = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix},$$

perfectly uncoupling in both Algorithm 3 and Algorithm 4 the prox updates for G into ones for G_1 and G_2. The condition (F^*-pcr) with $\bar{\rho} = \infty$ is then immediate from Lemma 2.1. Moreover, the Sobolev–Korn inequality [31] allows us to bound on a connected domain $\Omega \subset \mathbb{R}^2$ an optimal $\hat{w}$ to (77) as

$$\inf_{\bar{w} \text{ affine}} \|\hat{w} - \bar{w}\|_{L^2} \le C_\Omega \|\mathcal{E}\hat{w}\|_1 \le C_\Omega G_0(f)$$

for some constant $C_\Omega > 0$. We may assume that $\bar{w} = 0$, as the affine part of w is not used in (77). Therefore we may again replace $G_2 = 0$ by the artificial constraint $G_2(w) = \iota_{\|\cdot\|_{L^2} \le M}(w)$. By Lemma 2.1, G will then satisfy (G-pcr) with $\bar{\gamma}^\perp = \infty$.

5.3 Numerical Results

We demonstrate our algorithms on TGV^2 denoising and TV deblurring. Our tests are done on the photographs in Fig. 1, both at the original resolution of 768×512, and scaled down by a factor of 0.25 to 192×128 pixels. It is image #23 from the free Kodak image suite. Other images from the collection that we have experimented on give analogous computational results. For both of our example problems, we calculate a target solution by taking one million iterations of the basic PDHGM (3). We also tried interior point methods for this, but they are only practical for the smaller denoising problem.

We evaluate Algorithms 3 and 4 against the standard unaccelerated PDHGM of [3], as well as (a) the mixed-rate method of [5], denoted here C-L-O, (b) the relaxed PDHGM of [20,32], denoted here 'Relax', and (c) the adaptive PDHGM of [33], denoted here 'Adapt'. All of these methods are very closely linked and have comparable low costs for each step. This makes them straightforward to compare.

As we have discussed, for comparison and stopping purposes, we need to calculate a pseudo-duality gap as in [28], because the real duality gap is in practise infinite when A has a non-trivial nullspace. We do this dynamically; upgrading, the M in (76) every time, we compute the duality gap. For both of our example problems, we use for simplicity

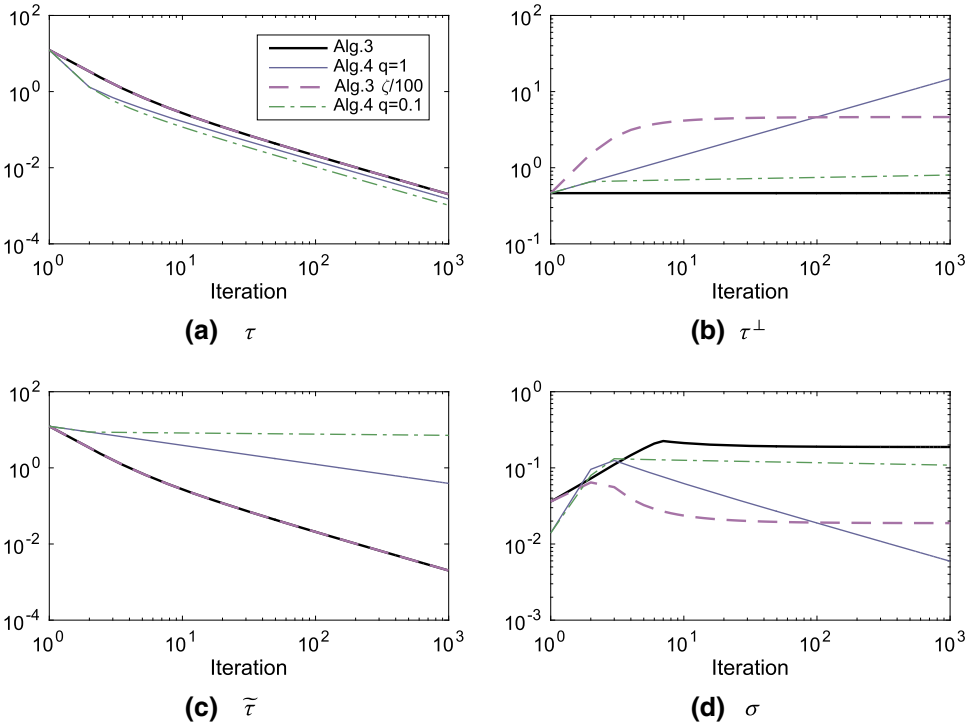

Fig. 2 Step length parameter evolution, both axes logarithmic. 'Alg.3' and 'Alg.4 q=1' have the same parameters as our numerical experiments for the respective algorithms, in particular $\zeta = \tau_0^{\perp,-2}$ for Algorithm 3, which yields constant $\tau^\perp$. 'Alg.3 $\zeta/100$' uses the value $\zeta = \tau_0^{\perp,-2}/100$, which causes $\tau^\perp$ to increase for some iterations. 'Alg.4 q=0.1' uses the value $q = 0.1$ for Algorithm 4, everything else being kept equal

$Z = L^2$ in (76). In the calculation of the final duality gaps comparing each algorithm, we then take as M the maximum over all evaluations of all the algorithms. This makes the results fully comparable. We always report the duality gap in decibels $10\log_{10}(\text{gap}^2/\text{gap}_0^2)$ relative to the initial iterate. Similarly, we report the distance to the target solution $\hat{u}$ in decibels $10\log_{10}(\|u^i - \hat{u}\|^2/\|\hat{u}\|^2)$, and the primal objective value $\text{val}(x) := G(x) + F(Kx)$ relative to the target as $10\log_{10}(\text{val}(x)^2/\text{val}(\hat{x})^2)$. Our computations were performed in MATLAB+C-MEX on a MacBook Pro with 16GB RAM and a 2.8 GHz Intel Core i5 CPU.

TGV^2 *denoising* The noise in our high-resolution test image, with values in the range $[0, 255]$, has standard deviation 29.6 or 12 dB. In the downscaled image, these become, respectively, 6.15 or 25.7 dB. As parameters (β, α) of the TGV^2 regularisation functional, we choose (4.4, 4) for the downscale image, and translate this to the original image by multiplying by the scaling vector $(0.25^{-2}, 0.25^{-1})$ corresponding to the 0.25 downscaling factor. See [34] for a discussion about rescaling and regularisation factors, as well as for a justification of the β/α ratio.

For the PDHGM and our algorithms, we take $\gamma = 0.5$, corresponding to the gap convergence results. We choose $\delta = 0.01$, and parametrise the PDHGM with $\sigma_0 = 1.9/\|K\|$ and $\tau_0^* = \tau_0 \approx 0.52/\|K\|$ solved from $\tau_0\sigma_0 = (1-\delta)\|K\|^2$. These are values that typically work well. For forward-differences discretisation of TGV^2 with cell width $h = 1$, we have $\|K\|^2 \leq 11.4$ [28]. We use the same value of δ for Algorithm 3 and Algorithm 4, but choose $\tau_0^\perp = 3\tau_0^*$, and $\tau_0 = \widetilde{\tau}_0 = 80\tau_0^*$. We also take $\zeta = \tau_0^{\perp,-2}$ for Algorithm 3. These values have been found to work well by trial and error, while keeping δ comparable to the PDHGM. A similar choice of τ_0 with a corresponding modification of σ_0 would significantly reduce the performance of the PDHGM. For Algorithm 4, we take exponent $q = 0.1$ for the sequence $\{a_i\}$. This gives in principle a mixed $O(1/N^{1.5}) + O(1/N^{0.5})$ rate, possibly improved by the convergence of the dual sequence. We plot the evolution of the step length for these and some other choices in Fig. 2. For the C-L-O, we use the detailed parametrisation from [35, Corollary 2.4], taking as Ω_Y the true L^2-norm Bregman divergence of $B(0,\alpha) \times B(0,\beta)$, and $\Omega_X = 10 \cdot \|f\|^2/2$ as a conservative estimate of a ball containing the true solution. For 'Adapt', we use the exact choices of α_0, η, and c from [33]. For 'Relax', we use the value 1.5 for the inertial ρ parameter of [32]. For both of these algorithms, we use the same choices of σ_0 and τ_0 as for the PDHGM.

We take fixed 20,000 iterations and initialise each algorithm with $y^0 = 0$ and $x^0 = 0$. To reduce computational overheads, we compute the duality gap and distance to target only every 10 iterations instead of at each iteration. The results are in Fig. 3 and Table 1. As we can see, Algorithm 3 performs extremely well for the low-resolution image, especially in its initial iterations. After about 700 or 200 iterations, depending on the criterion, the standard and relaxed PDHGM start to overtake. This is a general effect that we have seen in our tests: the standard PDHGM performs in practise very well asymptotically, although in principle all that exists is a $O(1/N)$ rate on the ergodic

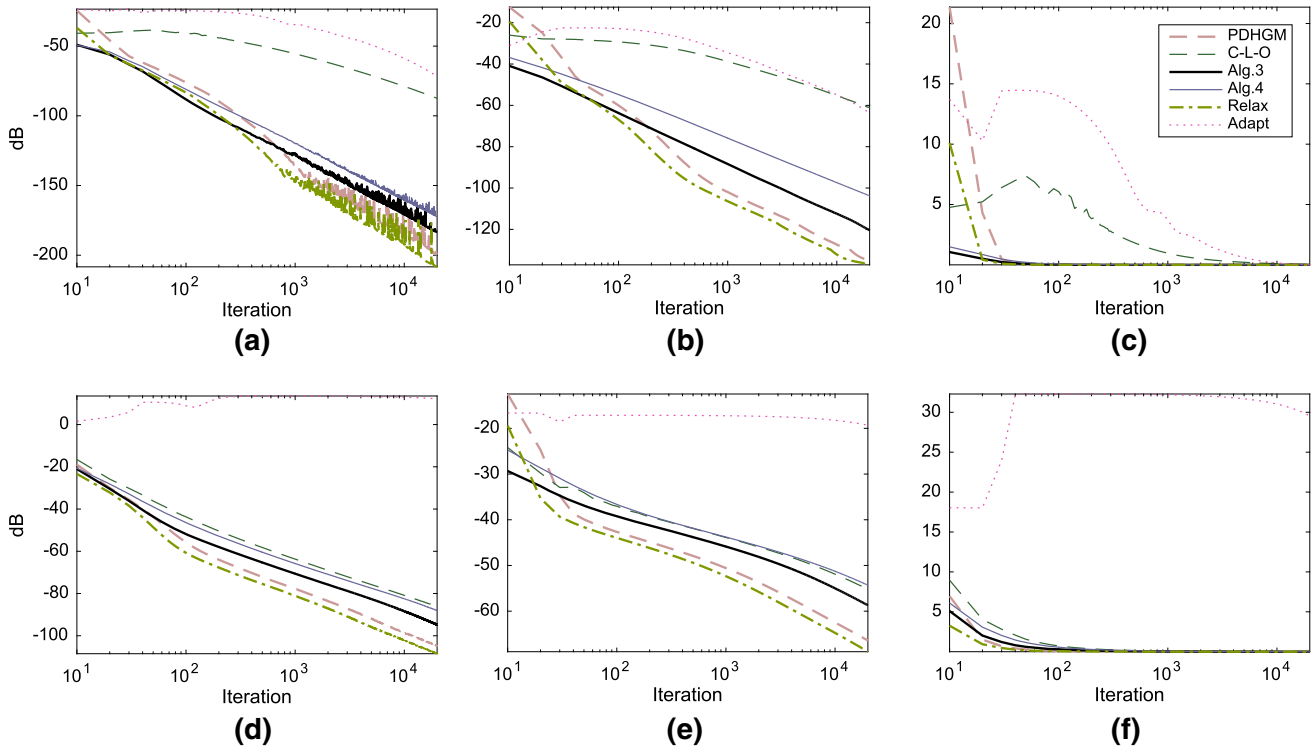

Fig. 3 TGV^2 denoising performance, 20,000 iterations, high- and low-resolution images. The plot is logarithmic, with the decibels calculated as in Sect. 5.3. The poor high-resolution results for 'Adapt' [33] have been omitted to avoid poor scaling of the plots. **a** Gap, low resolution, **b** target, low resolution, **c** value, low resolution, **d** gap, high resolution, **e** target, high resolution, **f** value, high resolution

Table 1 TGV^2 denoising performance, maximum 20,000 iterations

Low resolution						
Method	Gap ≤ -50 dB		Tgt ≤ -40 dB		Val ≤ 1 dB	
	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)
PDHGM	30	0.40	40	0.46	30	0.40
C-L-O	500	4.67	1210	11.31	970	9.04
Alg.3	20	0.29	10	0.22	20	0.29
Alg.4	20	0.47	20	0.47	20	0.47
Relax	20	0.34	30	0.45	20	0.34
Adapt	5360	106.63	2040	41.38	3530	70.78
High resolution						
Method	Gap ≤ -40 dB		Tgt ≤ -30 dB		Val ≤ 1 dB	
	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)
PDHGM	50	8.85	30	5.13	30	5.13
C-L-O	80	15.76	30	5.97	80	15.76
Alg.3	40	6.20	20	3.10	40	6.20
Alg.4	60	9.18	30	4.53	60	9.18
Relax	40	7.45	20	3.70	20	3.70
Adapt	–	–	–	–	–	–

The CPU time and number of iterations (at a resolution of 10) needed to reach given solution quality in terms of the duality gap, distance to target, or primal objective value

duality gap. Algorithm 4, by contrast, does not perform asymptotically so well. It can be extremely fast on its initial iterations, but then quickly flattens out. The C-L-O surprisingly performs better on the high-resolution image than on the low-resolution image, where it does somewhat poorly in comparison with the other algorithms. The adaptive PDHGM performs very poorly for TGV^2 denoising, and we have indeed excluded the high-resolution results from our reports to keep the scaling of the plots informative. Overall, Algorithm 3 gives good results fast, although the basic and relaxed PDHGM seems to perform, in practise, better asymptotically.

TV *deblurring* Our test image has now been distorted by Gaussian blur of standard deviation 4, which we intent to remove. We denote by $\hat{a}$ the Fourier presentation of the blur operator as discussed in Sect. 5.1. For numerical stability of the pseudo-duality gap, we zero out small entries, replacing this $\hat{a}$ by $\hat{a}\chi_{|\hat{a}(\cdot)|\geq\|\hat{a}\|_\infty/1000}(\xi)$. Note that this is only needed for the stable computation of G^* for the pseudo-duality gap, to compare the algorithms; the algorithms themselves are stable without this modification. To construct the projection operator P, we then set $\hat{p}(\xi) = \chi_{|\hat{a}(\cdot)|\geq 0.3\|\hat{a}\|_\infty}(\xi)$, and $P = \mathfrak{F}^*\hat{p}\mathfrak{F}$.

We use TV parameter 2.55 for the high-resolution image and the scaled parameter $2.55 * 0.15$ for the low-resolution image. We parametrise all the algorithms almost exactly as

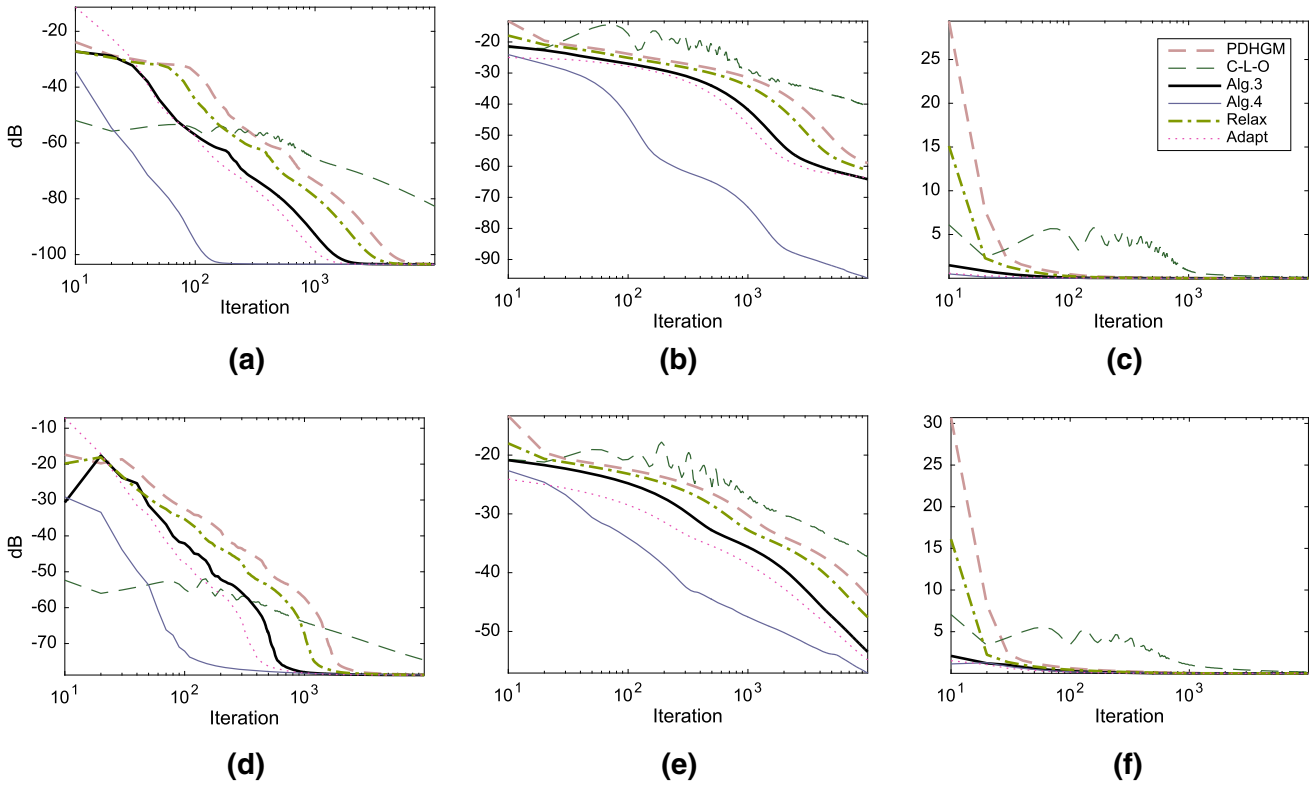

Fig. 4 TV deblurring performance, 10,000 iterations, high- and low-resolution images. The plot is logarithmic, with the decibels calculated as in Sect. 5.3. **a** Gap, low resolution. **b** Target, low resolution. **c** Value, low resolution. **d** Gap, high resolution. **e** Target, high resolution. **f** Value, high resolution

Table 2 TV deblurring performance, maximum 10,000 iterations

Method	Low resolution						High resolution					
	Gap ≤ -60 dB		Tgt ≤ -40 dB		Val ≤ 1 dB		Gap ≤ -60 dB		Tgt ≤ -30 dB		Val ≤ 1 dB	
	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)	Iter	Time (s)
PDHGM	390	2.53	2630	17.41	60	0.47	1180	118.30	970	98.98	70	6.59
C-L-O	600	3.81	8930	54.20	950	5.95	500	48.44	1940	187.42	1000	96.60
Alg.3	130	1.14	880	7.22	20	0.25	400	58.42	320	46.16	40	6.13
Alg.4	30	0.47	90	0.97	10	0.29	60	7.97	50	6.66	30	3.98
Relax	260	1.62	1750	11.34	40	0.29	790	77.31	650	63.84	50	5.29
Adapt	110	1.12	660	5.94	10	0.16	260	39.39	150	23.30	30	4.72

The CPU time and number of iterations (at a resolution of 10) needed to reach given solution quality in terms of the duality gap, distance to target, or primal objective value

TGV^2 denoising above, of course with appropriate Ω_U and $\|K\|^2 \leq 8$ corresponding to $K = \nabla$ [36]. The only difference in parameterisation is that we take $q = 1$ instead of $q = 0.1$ for Algorithm 4.

The results are in Fig. 4 and Table 2. It does not appear numerically feasible to go significantly below -100 or -80 dB gap. Our guess is that this is due to the numerical inaccuracies of the fast Fourier transform implementation in MATLAB. The C-L-O performs very well judged by the duality gap, although the images themselves and the primal objective value appear to take a little bit longer to converge. The relaxed PDHGM is again slightly improved from the standard PDHGM. The adaptive PDHGM performs very well, slightly outperforming Algorithm 3, although not Algorithm 4. This time Algorithm 4 performs remarkably well.

6 Conclusion

To conclude, overall, our algorithms are very competitive within the class of proposed variants of the PDHGM. Within our analysis, we have, moreover, proposed very streamlined derivations of convergence rates for even the standard PDHGM, based on the proximal point formulation and the idea of testing. Interesting continuations of this study include whether the condition $\hat{T}_i K = K \widetilde{T}_i$ can reasonably be relaxed such that $\hat{T}_i$ and $\widetilde{T}_i$ would not have to be scalars, as well as the relation to block coordinate descent methods, in particular [14,37].

Acknowledgements This research was started while T. Valkonen was at the Center for Mathematical Modeling at Escuela Politécnica Nacional in Quito, supported by a Prometeo scholarship of the Senescyt (Ecuadorian Ministry of Science, Technology, Education, and Innovation). In Cambridge, T. Valkonen has been supported by the EPSRC Grant EP/M00483X/1 "Efficient computational tools for inverse imaging problems". Thomas Pock is supported by the European Research Council under the Horizon 2020 programme, ERC starting Grant Agreement 640156.

Compliance with ethical standards

A Data Statement for the EPSRC This is primarily a theory paper, with some demonstrations on a photograph freely available from the Internet. As this article was written, the used photograph from the Kodak image suite was, in particular, available at http://r0k.us/graphics/kodak/. It has also been archived with our implementations of the algorithms at https://www.repository.cam.ac.uk/handle/1810/253697.

References

1. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1972)
2. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. SIAM (1999)
3. 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
4. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal–dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010). doi:10.1137/09076934X
5. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal–dual methods for a class of saddle point problems. SIAM J. Optim. **24**(4), 1779–1814 (2014). doi:10.1137/130919362
6. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005). doi:10.1007/s10107-004-0552-5
7. Beck, A., Teboulle, M.: Smoothing and first order methods: a unified framework. SIAM J. Optim. **22**(2), 557–580 (2012). doi:10.1137/100818327
8. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2015). doi:10.1007/s10208-013-9150-3
9. Beck, A., Teboulle, M.: A fast dual proximal gradient algorithm for convex minimization and applications. Oper. Res. Lett. **42**(1), 1–6 (2014). doi:10.1016/j.orl.2013.10.007
10. Valkonen, T.: A primal–dual hybrid gradient method for non-linear operators with applications to MRI. Inverse Probl. **30**(5), 055,012 (2014). doi:10.1088/0266-5611/30/5/055012
11. Benning, M., Knoll, F., Schönlieb, C.B., Valkonen, T.: Preconditioned ADMM with nonlinear operator constraint (2015). arXiv:1511.00425
12. Möllenhoff, T., Strekalovskiy, E., Moeller, M., Cremers, D.: The primal–dual hybrid gradient method for semiconvex splittings. SIAM J. Imaging Sci. **8**(2), 827–857 (2015). doi:10.1137/140976601
13. Lorenz, D., Pock, T.: An inertial forward-backward algorithm for monotone inclusions. J. Math. Imaging Vis. **51**(2), 311–325 (2015). doi:10.1007/s10851-014-0523-2
14. Fercoq, O., Bianchi, P.: A coordinate descent primal–dual algorithm with large step size and possibly non separable functions (2015). arXiv:1508.04625
15. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009). doi:10.1137/080716542
16. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**(11), 2419–2434 (2009). doi:10.1109/TIP.2009.2028250
17. Setzer, S.: Operator splittings, Bregman methods and frame shrinkage in image processing. Int. J. Comput. Vis. **92**(3), 265–280 (2011). doi:10.1007/s11263-010-0357-3
18. Valkonen, T.: Optimising big images. In: A. Emrouznejad (ed.) Big Data Optimization: Recent Developments and Challenges, Studies in Big Data, pp. 97–131. Springer, Berlin (2016). doi:10.1007/978-3-319-30265-2_5
19. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, Berlin (1998). doi:10.1007/978-3-642-02431-3
20. He, B., Yuan, X.: Convergence analysis of primal–dual algorithms for a saddle-point problem: from contraction perspective. SIAM J. Imaging Sci. **5**(1), 119–149 (2012). doi:10.1137/100814494
21. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal–dual algorithms in convex optimization. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 1762–1769 (2011). doi:10.1109/ICCV.2011.6126441
22. Rudin, W.: Functional Analysis. International series in Pure and Applied Mathematics. McGraw-Hill, New York (2006)
23. Bauschke, H., Combettes, P.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, Berlin (2011)
24. Hohage, T., Homann, C.: A generalization of the Chambolle–Pock algorithm to Banach spaces with applications to inverse problems (2014). arXiv:1412.0126
25. Chan, T., Shen, J.: Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods. Society for Industrial and Applied Mathematics (SIAM) (2005)
26. Benning, M., Gladden, L., Holland, D., Schönlieb, C.B., Valkonen, T.: Phase reconstruction from velocity-encoded MRI measurements—a survey of sparsity-promoting variational approaches. J. Magn. Reson. **238**, 26–43 (2014). doi:10.1016/j.jmr.2013.10.003

27. Ambrosio, L., Fusco, N., Pallara, D.: Functions of Bounded Variation and Free Discontinuity Problems. Oxford University Press, Oxford (2000)
28. Valkonen, T., Bredies, K., Knoll, F.: Total generalised variation in diffusion tensor imaging. SIAM J. Imaging Sci. **6**(1), 487–525 (2013). doi:10.1137/120867172
29. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**, 492–526 (2011). doi:10.1137/090769521
30. Bredies, K., Valkonen, T.: Inverse problems with second-order total generalized variation constraints. In: Proceedings of the 9th International Conference on Sampling Theory and Applications (SampTA) 2011, Singapore (2011)
31. Temam, R.: Mathematical Problems in Plasticity. Gauthier-Villars (1985)
32. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. (2015). doi:10.1007/s10107-015-0957-3
33. Goldstein, T., Li, M., Yuan, X.: Adaptive primal–dual splitting methods for statistical learning and image processing. Adv. Neural Inf. Process. Syst. **28**, 2080–2088 (2015)
34. de Los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel parameter learning for higher-order total variation regularisation models. J. Math. Imaging Vis. (2016). doi:10.1007/s10851-016-0662-8. Published online
35. Chen, K., Lorenz, D.A.: Image sequence interpolation using optimal control. J. Math. Imaging Vis. **41**, 222–238 (2011). doi:10.1007/s10851-011-0274-2
36. Chambolle, A.: An algorithm for mean curvature motion. Interfaces Free Bound. **6**(2), 195 (2004)
37. Suzuki, T.: Stochastic dual coordinate ascent with alternating direction multiplier method (2013). arXiv:1311.0622v1

Tuomo Valkonen received his Ph.D. from the University of Jyväskylä in 2008. Since then he has worked as researcher in Graz, Cambridge, and Quito. In February 2016 he started as a lecturer at the University of Liverpool.

Thomas Pock born 1978 in Graz, received his M.Sc. (1998–2004) and his Ph.D. (2005–2008) in Computer Engineering (Telematik) from Graz University of Technology. After a Postdoc position at the University of Bonn, he moved back to Graz University of Technology where he has been an Assistant Professor at the Institute for Computer Graphics and Vision. In 2013 he received the START price of the Austrian Science Fund (FWF) and the German Pattern recognition award of the German association for pattern recognition (DAGM) and in 2014, he received an starting grant from the European Research Council (ERC). Since June 2014, he is a Professor of Computer Science at Graz University of Technology (AIT Stiftungsprofessur "Mobile Computer Vision") and a principal scientist at the Digital Department of Safety and Security at the Austrian Institute of Technology (AIT). The focus of his research is the development of mathematical models for computer vision and image processing as well as the development of efficient algorithms to solve these models.