Acceleration of the PDHGM on strongly convex subspaces

Tuomo Valkonen[*]   **Thomas Pock**[†]

November 20, 2015

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.

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 minimax problem

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

One possibility is the primal-dual algorithm of Chambolle and Pock [11], a type of proximal point or extragradient method, also classified as the "modified primal-dual hybrid gradient method" or PDHGM by Esser [17]. If either G of F^* is strongly convex, the method can be accelerated to produce Nesterov's [23] optimal $O(1/N^2)$ rates. 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 the

---

[*]   Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom. tuomo.valkonen@iki.fi

[†]   Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria. Digital Safety & Security Department, AIT Austrian Institute of Technology GmbH, 1220 Vienna, Austria. pock@icg.tugraz.at

data sciences, as we will more closely review in Section 5. Under such *partial strong convexity*, can we obtain a method that would give an accelerated rate of convergence at least for Px?

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, our results are similar to the "optimal" algorithm of Chen et al. [14]. 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 [11].

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

$$\langle \partial G(x') - \partial G(x), \widetilde{T}^{-1}(x' - x) \rangle \geq \|x' - x\|^2_{\widetilde{T}^{-1,*}\Gamma'} - \psi_{\widetilde{T}^{-1,*}(\Gamma'-\Gamma)}(x' - x), \qquad (1.1)$$

the full details of which we provide in Section 2. In this, we make the monotonicity dependent on the step length operator $\widetilde{T}$. Secondly, our factor of strong convexity is the operator Γ, which is however shifted in (1.1) into a *penalty term* ψ through the introduction of additional strong monotonicity in terms of $\Gamma' \geq \Gamma$. This exact procedure can be seen as a type of smoothing, famously studied by Nesterov [24], and more recently, for instance, by Beck and Teboulle [4]. In these approaches, one computes *a priori* a level of smoothing—comparable to Γ'—needed to achieve certain quality of solution, and then solves a smoothed problem at the optimal $O(1/N^2)$ rate. However, to achieve a better solution than the a priori chosen quality, one needs to solve a new problem from scratch, or to develop restarting strategies. Our approach does not depend on restarting and a priori chosen solution qualities. Indeed, Γ' is controlled automatically. In most applications, $\psi_{\widetilde{T}^{-1,*}(\Gamma'-\Gamma)}(x' - x) = \widetilde{\tau}^{-1}\gamma^\perp C$ for $\gamma^\perp$ the introduced strong monotonicity on the orthogonal complement $X_0^\perp$. This kind of constant ψ can in particular be achieved on bounded domains, as was also employed for the aforementioned mixed-rate algorithm [14].

The "fast dual proximal gradient method", or FDPG [5], 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 form the PDHGM, such as variants for non-linear K [30, 7], and non-convex G [22]. It has been the basis for inertial methods for monotone inclusions [21], and primal-dual stochastic coordinate descent methods without separability requirements [18]. Finally, the FISTA [3, 2] can be seen as a primal-only version of the PDHGM. Not attempting to do full justice here to the large family of closely-related methods, we point to [17, 28, 31] 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 Section 2.

A byproduct of this work is the shortest convergence rate proof for the accelerated PDHGM known to us. Secondly, in Section 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 do some of the groundwork for these algorithms in Section 3. We finish the study with numerical experiments in Section 5. The main results of interest for readers wishing to apply our work are Algorithms 3 and 4.

2   A general primal-dual method

2.1   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 $\lim_{\|x\| \to \infty} G(x) = \infty$ and $\lim_{\|y\| \to \infty} F^*(y) = \infty$ [16, Proposition VI.1.2 & Proposition VI.2.2]. More generally the existence has to be proved explicitly. In finite dimensions, see, e.g., [26] for sufficient conditions.

The primal-dual method of Chambolle and Pock [11] for (P) consists of iterating the system

$$x^{i+1} := (I + \tau_i \partial G)^{-1}(x^i - \tau_i K^* y^i), \tag{2.1a}$$

$$\bar{x}^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \tag{2.1b}$$

$$y^{i+1} := (I + \sigma_{i+1} \partial F^*)^{-1}(y^i + \sigma_{i+1} K \bar{x}^{i+1}). \tag{2.1c}$$

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. If G is strongly convex with factor γ, we may accelerate

$$\omega_i := 1/\sqrt{1 + 2\gamma \tau_i}, \quad \tau_{i+1} := \tau_i \omega_i, \quad \text{and} \quad \sigma_{i+1} := \sigma_i/\omega_i, \tag{2.2}$$

to achieve $O(1/N^2)$ convergence rates. To motivate our choices later on, observe that σ_0 is never needed, if we equivalently parametrise the algorithm by $\delta = 1 - \|K\|^2 \tau_0 \sigma_0 > 0$, which gives the fixed ratio of σ_i to τ_i.

We note that the order of the steps in (2.1) is reversed from the original ordering in [11]. This is because with the present order, the method (2.1) may also be written in

the proximal point form. This formulation, first observed in [19] and later utilised in [30, 25, 7], is also what we will use to streamline our analysis. Introducing the general variable splitting notation,

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

the system (2.1) then reduces into

$$0 \in H(u^{i+1}) + M_{\text{basic}, i}(u^{i+1} - u^i), \tag{2.3}$$

for the monotone operator

$$H(u) := \begin{pmatrix} \partial G(x) + K^* y \\ \partial F^*(y) - Kx \end{pmatrix}, \tag{2.4}$$

and the *preconditioning* or *step-length operator*

$$M_{\text{basic}, i} := \begin{pmatrix} 1/\tau_i & -K^* \\ -\omega_i K & 1/\sigma_{i+1} \end{pmatrix}. \tag{2.5}$$

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

2.2 Abstract partial monotonicity

Our plan now is to formulate a general version of (2.1), 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 (2.1b) and (2.5), operating as $\hat{T}_{i+1} K \widetilde{T}_i^{-1} \approx \omega_i K$ from both sides of K.

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 $T \in \mathcal{L}(X; X)$, the notation $T \geq 0$ means that T is positive semidefinite. In this case, we also denote

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

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

$$\langle a, b \rangle_M := \langle Ma, b \rangle, \quad \text{and} \quad \|a\|_M := \sqrt{\langle a, a \rangle_M}. \tag{2.6}$$

To start the algorithm derivation, we now formulate abstract forms of partial strong monotonicity. As a first step, we take subspaces 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

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

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 $\widetilde{T}_1, \widetilde{T}_2 \in \widetilde{\mathcal{T}}$.

4

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*, which we take to mean that

$$\langle \partial G(x') - \partial G(x), \widetilde{T}^{-1}(x'-x) \rangle \geq \|x'-x\|^2_{\widetilde{T}^{-1,*}\Gamma'} - \psi_{\widetilde{T}^{-1,*}(\Gamma'-\Gamma)}(x'-x),$$
$$(x, x' \in X;\ \widetilde{T} \in \widetilde{\mathcal{T}};\ \Gamma' \in [0, \Gamma] + \widetilde{\mathcal{K}}). \quad \text{(G-PM)}$$

for some family of functionals $\{\psi_T : X \to \mathbb{R}\}$, and a linear operator $0 \leq \Gamma \in \mathcal{L}(X; X)$ which models partial strong monotonicity. 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 Section 2.3.

Example 2.1. $G(x) = \|f - Ax\|^2$ satisfies (G-PM) with $\Gamma = A^*A$, $\widetilde{T} = \{0\}$, and $\psi \equiv 0$ for any invertible $\widetilde{T}$.

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

$$\langle \partial F^*(y') - \partial F^*(y), \hat{T}^{-1,*}(y'-y) \rangle \geq \|y'-y\|^2_{\hat{T}^{-1}R} - \phi_{\hat{T}^{-1}R}(y'-y),$$
$$(y, y' \in Y;\ \hat{T} \in \hat{\mathcal{T}};\ R \in \hat{\mathcal{K}}) \quad \text{(F*-PM)}$$

for some family of functionals $\{\phi_T : Y \to \mathbb{R}\}$.

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. As discussed in the introduction, these functionals can be seen as an abstract approach to smoothing, however without any "restarting" requirements on the algorithm. With this, $\widetilde{\mathcal{K}}$ and $\hat{\mathcal{K}}$ can be seen as sets of admissible smoothing parameters.

The next lemma demonstrates how the existence of ϕ or ψ might be satisfied when all the parameters are scalar.

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

$$G(x') - G(x) \geq \langle \partial G(x), x'-x \rangle + \frac{\gamma}{2}\left(\|x'-x\|^2 - C_\psi\right), \quad (2.7)$$

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 (2.7) holds irrespective of γ and C. If $x \notin A$, we have $\partial G(x) = \emptyset$, so (2.7) 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, (2.7) holds. $\quad\square$

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

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 $0 \le \Gamma \in \mathcal{L}(X; X)$. 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

$$S_j M_j \ge \delta \begin{pmatrix} \widetilde{T}_j^{-1,*} T_j^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.8}$$

1: Choose initial iterates $x^0 \in X$ and $y^0 \in Y$.
2: **repeat**
3:     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 (2.8) with $j = i + 1$, as well as the condition

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

    for some $0 \le R_{i+1} \in \hat{\mathcal{K}}$ and $\Gamma_i \in [0, \Gamma] + \widetilde{\mathcal{K}}$.
4:     Perform the updates

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

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

2.3 A general algorithm and the idea of testing

The only change we make to the proximal point formulation (2.3) of the method (2.1), is to replace the basic step length or preconditioning operator $M_{\text{basic}, i}$ by the operator

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

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*

$$S_i := \begin{pmatrix} \widetilde{T}_i^{-1,*} & 0 \\ 0 & \hat{T}_{i+1}^{-1} \end{pmatrix}. \tag{2.9}$$

It will turn out that multiplying or "testing" (PP) by this operator will allow us to derive convergence rates. Hence the terminology. The testing of (PP) by S_i is also why we introduced testing into the monotonicity conditions (G-PM) and (F*-PM).

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 $R_{i+1} \in \hat{\mathcal{K}}$. With this, taking a fixed $\delta > 0$, the properties

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

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

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

2.4 A simplified condition

We expand

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

as well as

$$S_i\bar{\Gamma}_i = \begin{pmatrix} 2\widetilde{T}_i^{-1,*}\Gamma_i & \widetilde{T}_i^{-1,*}K^* - K^*\hat{T}_{i+1}^{-1,*} \\ K\widetilde{T}_i^{-1} - \hat{T}_{i+1}^{-1}K & 2\hat{T}_{i+1}^{-1}R_{i+1} \end{pmatrix}, \tag{2.10}$$

and

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

By Young's inequality, (C2) is thus satisfied when for some invertible $Z_i \in \mathcal{L}(X; X)$,

$$\hat{T}_{i+1}^{-1}\Sigma_{i+1}^{-1} \geq KZ_i^{-1}Z_i^{-1,*}K^*, \quad \text{and} \quad (1 - \delta)\widetilde{T}_i^{-1,*}T_i^{-1} \geq \widetilde{T}_i^{-1,*}Z_i^*Z_i\widetilde{T}^{-1}.$$

The second constraint is satisfied as an equality if

$$Z_i^*Z_i = (1 - \delta)T_i^{-1}\widetilde{T}_i. \tag{2.11}$$

Note that this choice will also be optimal for the first constraint. By the spectral theorem for self-adjoint operators on Hilbert spaces (e.g., [27, Chapter 12]), we can make the choice (2.11) if

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

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

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

Also, (C1) can be rewritten

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

2.5 Basic convergence result

Our main convergence result is as follows.

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 $R_{i+1} \in \hat{\mathcal{K}}$. Then the iterates of Algorithm 1 satisfy*

$$\frac{\delta}{2}\|x^N - \widehat{x}\|^2_{\widetilde{T}_N^{-1,*}T_N^{-1}} \leq C_0 + \sum_{i=0}^{N-1} D_{i+1}, \quad (N \geq 1), \tag{2.12}$$

for

$$D_{i+1} := \psi_{\widetilde{T}_i^{-1,*}(\Gamma_i - \Gamma)}(x^{i+1} - \widehat{x}) + \phi_{\hat{T}_{i+1}^{-1}R_{i+1}}(y^{i+1} - \widehat{y}), \quad and \quad C_0 := \frac{1}{2}\|u^0 - \widehat{u}\|^2_{S_0 M_0}. \tag{2.13}$$

Remark 2.1. The term D_{i+1} coming from the dissatisfaction of strong convexity, penalises the basic convergence rate. 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 $D_{i+1} \leq C$ 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 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 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 we can however usually ensure.

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

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

Recalling the definition of S_i from (2.9), and of H from (2.4), it follows

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

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

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

Using the expression (2.10) for $S_i \bar{\Gamma}_i$, and (2.13) for D_{i+1}, we thus deduce

$$\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \geq \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{S_i \bar{\Gamma}_i} - D_{i+1}. \tag{2.14}$$

8

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

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

That is

$$
\langle u^{i+1} - u^i, u^{i+1} - \widehat{u} \rangle_M = \frac{1}{2}\|u^{i+1} - u^i\|_M^2 - \frac{1}{2}\|u^i - \widehat{u}\|_M^2 + \frac{1}{2}\|u^{i+1} - \widehat{u}\|_M^2. \tag{2.15}
$$

Using this and (PP), we calculate

$$
\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u}) \rangle \ni -\frac{1}{2}\|u^{i+1} - u^i\|_{S_i M_i}^2 + \frac{1}{2}\|u^i - \widehat{u}\|_{S_i M_i}^2 - \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_i M_i}^2.
$$

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

$$
\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u}) \rangle \leq \frac{1}{2}\|u^i - \widehat{u}\|_{S_i M_i}^2 - \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_{i+1} M_{i+1}}^2 + \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_i \bar{\Gamma}_i}^2. \tag{2.16}
$$

Combining (2.14) with (2.16), we thus obtain

$$
\frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_{i+1} M_{i+1}}^2 \leq \frac{1}{2}\|u^i - \widehat{u}\|_{S_i M_i}^2 + D_{i+1}. \tag{2.17}
$$

Summing (2.17) over $i = 1, \dots, N - 1$, and applying (C2) to estimate

$$
\frac{\delta}{2}\|x^{i+1} - \widehat{x}\|_{\widetilde{T}_N^{-1,*} T_N^{-1}}^2 \leq \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_N M_N}^2,
$$

we obtain (2.12). $\qquad\qquad\square$

3 Scalar diagonal updates and the ergodic duality gap

One relatively easy way to satisfy (G-PM), (F*-PM), (C1) and (C2), is to take the "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 specialising now Theorem 2.1 to the scalar case. We then explore ways to add estimates of the ergodic duality gap into (2.12). 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 diagonal updates to derive estimates on the ergodic duality gap.

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 $0 \leq \Gamma \in \mathcal{L}(X;X)$. A choice of $\delta \in (0,1)$. Initial invertible step length operators $T_0 \in Q$ 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

$$\widetilde{\omega}_i(I + 2\Gamma_i T_i)\Omega_i \geq I, \quad \text{and} \quad T_i\Omega_i \in Q. \tag{3.1a}$$

4:      Set

$$T_{i+1} := T_i\Omega_i, \quad \text{and} \quad \widetilde{\tau}_{i+1} := \widetilde{\tau}_i\widetilde{\omega}_i. \tag{3.1b}$$

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

$$\Sigma_i^{-1} + 2R_i \geq \widetilde{\omega}_i^{-1}\Sigma_{i+1}^{-1} \geq (1-\delta)^{-1}KT_iK^*. \tag{3.1c}$$

6:      Perform the updates

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

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

3.1 Scalar specialisation of Algorithm 1

We take $\widetilde{\Omega}_i = \widetilde{\omega}_i I$, $\widetilde{T}_i = \widetilde{\tau}_i I$, and $\hat{T}_i = \widetilde{\tau}_i I$ for some $\widetilde{\omega}_i, \widetilde{\tau}_i > 0$. With this (C2′) becomes

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

while, the diagonal terms cancelling out, (C1′) becomes

$$\begin{aligned}
\widetilde{\tau}_i^{-1}(I + 2\Gamma_i T_i)T_i^{-1} &\geq \widetilde{\tau}_{i+1}^{-1}T_{i+1}^{-1}, \quad \text{and} \\
\widetilde{\tau}_{i+1}^{-1}(\Sigma_{i+1}^{-1} + 2R_{i+1}) &\geq \widetilde{\tau}_{i+2}^{-1}\Sigma_{i+2}^{-1}.
\end{aligned} \tag{C1″}$$

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

$$\psi_T(-x) = \psi_T(x), \quad \text{and} \quad \psi_{\alpha T}(x) = \alpha\psi_T(x), \quad (x \in X; 0 < \alpha \in \mathbb{R}). \tag{3.2}$$

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

$$\langle \partial G(x') - \partial G(x), x' - x \rangle \geq \|x' - x\|_{\Gamma'}^2 - \psi_{\Gamma'-\Gamma}(x' - x), \quad (x, x' \in X; \Gamma' \in [0,\Gamma] + \widetilde{\mathcal{K}}),$$
$$\tag{G-pm}$$

and

$$\langle \partial F^*(y') - \partial F^*(y), y' - y \rangle \geq \|y' - y\|_R^2 - \phi_R(y' - y), \quad (y, y' \in Y; R \in \hat{\mathcal{K}}).$$
(F*-pm)

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 Ω_i and $\widetilde{\omega}_i$ satisfying $T_{i+1} = T_i \Omega_i$ and $\widetilde{\tau}_{i+1} = \widetilde{\tau}_i \widetilde{\omega}_i$, we reorganise (C1″) and (C2″) into the parameter update rules (3.1) 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. Observe that $\widetilde{\tau}_i$ and $\hat{\tau}_i$ disappear from the algorithm aside from the residual factor $\widetilde{\omega}_i$, which can give different over-relaxation rates in the rule for $\bar{x}^{i+1}$ compared to ω_i in (2.1). Moreover, the parameter $\widetilde{\tau}_i$ will still play a critical role in our study of convergence rate estimates.

Example 3.1 (The method of Chambolle and Pock). Let G be strongly convex with factor $\gamma \geq 0$. 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

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

Updating σ_{i+1} such that the last inequality holds as an equality, we recover the accelerated PDHGM (2.1)+(2.2). 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

$$G(x') - G(x) \geq \langle \partial G(x), x' - x \rangle + \frac{1}{2}\|x' - x\|_{\Gamma'}^2 - \frac{1}{2}\psi_{\Gamma'-\Gamma}(x' - x),$$
$$(x, x' \in X; \Gamma' \in [0, \Gamma] + \widetilde{\mathcal{K}}). \quad \text{(G-pc)}$$

and

$$F^*(y') - F^*(y) \geq \langle \partial F^*(y), y' - y \rangle + \frac{1}{2}\|y' - y\|_R^2 - \frac{1}{2}\psi_R(y' - y),$$
$$(y, y' \in Y; R \in \hat{\mathcal{K}}). \quad \text{(F*-pc)}$$

It is easy to see that these imply (G-pm) and (F*-pm).

To define an ergodic duality gap, we set

$$\widetilde{q}_N := \sum_{j=0}^{N-1} \widetilde{\tau}_j^{-1}, \quad \text{and} \quad \hat{q}_N := \sum_{j=0}^{N-1} \hat{\tau}_{j+1}^{-1}, \tag{3.3}$$

and define the weighted averages

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

With these, the ergodic duality gap at iteration N is defined as

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

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 $T_i \in Q$, $\Sigma_i \in \mathcal{L}(Y;Y)$,*

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

as well as $\Gamma_i \in ([0,\Gamma] + \widetilde{\mathcal{K}})/2$ and $R_i \in \hat{\mathcal{K}}/2$. Then the iterates of Algorithm 2 satisfy

$$\frac{\delta}{2} \|x^N - \hat{x}\|_{\widetilde{\tau}_N^{-1} T_N^{-1}}^2 + \widetilde{q}_N \mathcal{G}^N \le C_0 + \sum_{i=0}^{N-1} D_{i+1}. \tag{3.4}$$

Here C_0 is as in (2.13), *and D_{i+1} simplifies into*

$$D_{i+1} = \widetilde{\tau}_i^{-1} \psi_{(\Gamma_i - \Gamma)}(x^{i+1} - \hat{x}) + \widetilde{\tau}_{i+1}^{-1} \phi_{R_{i+1}}(y^{i+1} - \hat{y}). \tag{3.5}$$

If only (G-pm) *and* (F*-pm) *hold instead of* (G-pc) *and* (F*-pc), *or we take $R_i \in \hat{\mathcal{K}}$ and $\Gamma_i \in [0,\Gamma] + \widetilde{\mathcal{K}}$, then* (3.4) *holds with $\mathcal{G}^N := 0$.*

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 $\hat{\tau}_j = \widetilde{\tau}_j = \tau_0$. Thus presently $\widetilde{q}_N = N\tau_0$.

Example 3.3 (Full acceleration). With $\gamma > 0$ in Example 3.1, we know from [11, Corollary 1] that

$$\lim_{N \to \infty} N\tau_N \gamma = 1. \tag{3.6}$$

Thus $\widetilde{q}_N$ is of the order N^2. So is $\widetilde{\tau}_N T_N = \tau_N^2 I$. Therefore, (3.4) 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 (2.2) to $\omega_i = 1/\sqrt{1 + \gamma\tau_i}$.

Remark 3.2. The result (3.6) can be improved to estimate $\tau_N \le C_\tau/N$ without a qualifier $N \ge N_0$. Indeed, from [11, Lemma 2] we know for the rule $\omega_i = 1/\sqrt{1 + 2\gamma\tau_i}$ that 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

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

In particular,

$$\tau_\ell \leq \frac{1}{\gamma\left(\frac{1}{\gamma\tau_0} + \frac{\ell}{1+\gamma\tau_0}\right)} = \frac{1 + \gamma\tau_0}{\tau_0^{-1} + \gamma\ell} \leq \frac{\gamma^{-1} + \tau_0}{\ell}.$$

Therefore, $\tau_N \leq C_\tau/N$ for $C_\tau := \gamma^{-1} + \tau_0$. Moreover, $\tau_N^{-1} \leq \tau_0^{-1} + \gamma N$.

Proof of Theorem 3.1. The final non-gap estimate is a direct consequence of Theorem 2.1, so we concentrate on the gap estimate. We begin by expanding

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

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

$$\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \geq \widetilde{\tau}_i^{-1}\left(G(x^{i+1}) - G(\widehat{x}) + \frac{1}{2}\|x^{i+1} - \widehat{x}\|_{2\Gamma_i}^2 - \frac{1}{2}\psi_{2\Gamma_i}(x^{i+1} - \widehat{x})\right)$$
$$+ \widehat{\tau}_{i+1}^{-1}\left(F^*(y^{i+1}) - F^*(\widehat{y}) + \frac{1}{2}\|y^{i+1} - \widehat{y}\|_{2R_{i+1}}^2 - \frac{1}{2}\phi_{2R_{i+1}}(y^{i+1} - \widehat{y})\right)$$
$$- \widetilde{\tau}_i^{-1}\langle y^{i+1}, K\widehat{x}\rangle + \widehat{\tau}_{i+1}^{-1}\langle \widehat{y}, Kx^{i+1}\rangle + (\widetilde{\tau}_i^{-1} - \widehat{\tau}_{i+1}^{-1})\langle y^{i+1}, Kx^{i+1}\rangle.$$

Using (2.6) and (3.2), we can make all of the factors "2" and "1/2" in this expression annihilate each other. With D_{i+1} as in (3.5) (equivalently (2.13)) we therefore have

$$\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \geq \widetilde{\tau}_i^{-1}\left(G(x^{i+1}) - G(\widehat{x}) + \langle \widehat{y}, Kx^{i+1}\rangle\right) + \|x^{i+1} - \widehat{x}\|_{\widetilde{\tau}_i^{-1}\Gamma_i}^2$$
$$+ \widehat{\tau}_{i+1}^{-1}\left(F^*(y^{i+1}) - F^*(\widehat{y}) - \langle y^{i+1}, K\widehat{x}\rangle\right) + \|y^{i+1} - \widehat{y}\|_{\widehat{\tau}_{i+1}^{-1}R_{i+1}}^2$$
$$+ (\widetilde{\tau}_i^{-1} - \widehat{\tau}_{i+1}^{-1})\left(\langle y^{i+1} - \widehat{y}, K(x^{i+1} - \widehat{x})\rangle - \langle \widehat{y}, K\widehat{x}\rangle\right) - D_{i+1}.$$

A little bit of reorganisation and referral to (2.10) for the expansion of $S_i\bar{\Gamma}_i$ thus gives

$$\langle H(u^{i+1}), S_i^*(u^{i+1} - \widehat{u})\rangle \geq \widetilde{\tau}_i^{-1}\left(G(x^{i+1}) - G(\widehat{x}) + \langle \widehat{y}, Kx^{i+1}\rangle\right)$$
$$+ \widehat{\tau}_{i+1}^{-1}\left(F^*(y^{i+1}) - F^*(\widehat{y}) - \langle y^{i+1}, K\widehat{x}\rangle\right) \qquad (3.7)$$
$$- (\widetilde{\tau}_i^{-1} - \widehat{\tau}_{i+1}^{-1})\langle \widehat{y}, K\widehat{x}\rangle + \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_i\bar{\Gamma}_i}^2 - D_{i+1}.$$

Let us write

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

13

Observing here the switches between the indices $i + 1$ and i of the step length parameters in comparison to the last step of (3.7), we thus obtain

$$\langle H(u^{i+1}), S_i(u^{i+1} - \widehat{u}) \rangle \geq \mathcal{G}_+^i(u^{i+1}, \widehat{u}) - \mathcal{G}_+^i(\widehat{u}, \widehat{u}) + \frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_i \widetilde{\Gamma}_i}^2 - D_{i+1}. \qquad (3.8)$$

Using (2.16), we now obtain

$$\frac{1}{2}\|u^{i+1} - \widehat{u}\|_{S_{i+1}M_{i+1}}^2 + \mathcal{G}_+^i(u^{i+1}, \widehat{u}) - \mathcal{G}_+^i(\widehat{u}, \widehat{u}) \leq \frac{1}{2}\|u^i - \widehat{u}\|_{S_i M_i}^2 + D_{i+1}.$$

Summing this for $i = 0, \ldots, N - 1$ gives with C_0 from (2.13) the estimate

$$\frac{1}{2}\|u^N - \widehat{u}\|_{S_N M_N}^2 + \sum_{i=0}^{N-1}\left(\mathcal{G}_+^i(u^{i+1}, \widehat{u}) - \mathcal{G}_+^i(\widehat{u}, \widehat{u})\right) \leq C_0 + \sum_{i=0}^{N-1} D_{i+1}. \qquad (3.9)$$

We want to estimate the sum of the gaps $\mathcal{G}_+^i$ in (3.9). Using the convexity of G and F^*, we observe

$$\sum_{i=0}^{N-1} \widetilde{\tau}_i^{-1} G(x^{i+1}) \geq \widetilde{q}_N G(x_N), \quad \text{and} \quad \sum_{i=0}^{N-1} \hat{\tau}_{i+1}^{-1} F^*(y^{i+1}) \geq \hat{q}_N F^*(y_N). \qquad (3.10)$$

Also, by (3.3) and simple reorganisation

$$\sum_{i=0}^{N-1} \widetilde{\tau}_{i+1}^{-1} G(\widehat{x}) = \widetilde{q}_N G(\widehat{x}) + \widetilde{\tau}_N^{-1} G(\widehat{x}) - \widetilde{\tau}_0^{-1} G(\widehat{x}), \quad \text{and} \qquad (3.11)$$

$$\sum_{i=0}^{N-1} \hat{\tau}_i^{-1} F^*(\widehat{y}) = \hat{q}_N F^*(y_N) - \hat{\tau}_N^{-1} F^*(\widehat{y}) + \hat{\tau}_0^{-1} F^*(\widehat{y}). \qquad (3.12)$$

All of (3.10)–(3.12) together give

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

Another use of (3.3) gives

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

Thus

$$\sum_{i=0}^{N-1}\left(\mathcal{G}_+^i(u^{i+1}, \widehat{u}) - \mathcal{G}_+^i(\widehat{u}, \widehat{u})\right) \geq \widetilde{q}_N \mathcal{G}^N + r_N, \qquad (3.13)$$

14

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 $\hat{u} = (\hat{x}, \hat{y})$ to (OC), $K\hat{x} \in \partial F^*(\hat{y})$, so $r_N \geq 0$ 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 (3.13) in (3.9) therefore gives

$$\frac{1}{2}\|u^N - \hat{u}\|^2_{S_N M_N} + \widetilde{q}_N \mathcal{G}^N + r_N \leq C_0 + \sum_{i=0}^{N-1} D_{i+1}. \tag{3.14}$$

A referral to (C2) to estimate $S_N M_N$ from below shows (3.4), concluding the proof. □

Remark 3.3. We only used the assumption $\widetilde{\tau}_i = \hat{\tau}_i$ to bound $r_N \geq 0$. It is possible to streamline the proof if in addition to this we assume $\{\widetilde{\tau}_i\}$ to be non-increasing instead of merely satisfying (C3''), and define y^N based on $\widetilde{q}^N$ instead of $\hat{q}^N$.

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 Section 3, this is easiest when for some $\widetilde{\tau}_i > 0$ we set

$$\widetilde{T}_i = \widetilde{\tau}_i I, \quad \text{and} \quad \hat{T}_i = \widetilde{\tau}_i I. \tag{4.1}$$

The result 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, 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

$$\mathcal{F}(K T_i K^*) \geq K T_i K^*, \tag{4.2}$$

our approach is to define

$$\Sigma_{i+1}^{-1} := \widetilde{\omega}_i (1-\delta)^{-1} \mathcal{F}(K T_i K^*). \tag{4.3}$$

Then (C2'') is satisfied provided $T_i^{-1} \in Q$. 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

$$\widetilde{\tau}_i^{-1}(I + 2\Gamma T_i) T_i^{-1} - \widetilde{\tau}_{i+1}^{-1} T_{i+1}^{-1} \geq -2\widetilde{\tau}_i^{-1}(\Gamma_i - \Gamma), \quad \text{and} \tag{4.4a}$$

$$\frac{1}{1-\delta}\left(\widetilde{\tau}_i^{-1} \mathcal{F}(K T_i K^*) - \widetilde{\tau}_{i+1}^{-1} \mathcal{F}(K T_{i+1} K^*)\right) \geq -2\widetilde{\tau}_{i+1}^{-1} R_{i+1}. \tag{4.4b}$$

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

4.2 When Γ is 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, $P^2 = I$, and self-adjoint, $P^* = P$. We let $P^\perp := I - P$. Then $P^\perp P = PP^\perp = 0$. With this, we assume for some $\bar{\gamma}^\perp > 0$ that

$$[0, \bar{\gamma}^\perp P^\perp] \subset \widetilde{\mathcal{K}}. \tag{4.5}$$

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 $0 \le \gamma \le \lambda\bar{\gamma}$, and $0 \le \gamma_i^\perp \le \lambda\bar{\gamma}^\perp$, and set

$$T_i = \tau_i P + \tau_i^\perp P^\perp, \quad \Omega_i = \omega_i P + \omega_i^\perp P^\perp, \quad \text{and} \quad \Gamma_i = \gamma P + \gamma_i^\perp P^\perp. \tag{4.6}$$

With this, $\tau_i, \tau_i^\perp > 0$ guarantee $T_i \in Q$. 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

$$KT_iK^* = \tau_i KPK^* + \tau_i^\perp KP^\perp K^* = (\tau_i - \tau_i^\perp)KPK^* + \tau_i^\perp KK^*,$$

we are encouraged to take

$$\mathcal{F}(KT_iK^*) := \max\{0, \tau_i - \tau_i^\perp\}\|KP\|^2 I + \tau_i^\perp\|K\|^2 I. \tag{4.7}$$

Remark 4.1. If we required $\tau_i^\perp \ge \tau_i$, a simpler choice would be $\mathcal{F}(KT_iK^*) = \tau_i^\perp\|K\|^2 I$. Numerical experiments however suggest $\tau_0^\perp \ll \tau_0$ being beneficial. Nevertheless, for large enough i, the condition $\tau_i^\perp \ge \tau_i$ will hold in our algorithms.

Observe that (4.7) satisfies (4.2). Inserting (4.7) into (4.3), we obtain

$$\Sigma_{i+1} = \sigma_{i+1}I \quad \text{with} \quad \sigma_{i+1}^{-1} = \frac{\widetilde{\omega}_i}{1 - \delta}\left(\max\{0, \tau_i - \tau_i^\perp\}\|KP\|^2 + \tau_i^\perp\|K\|^2\right). \tag{4.8}$$

Since Σ_{i+1} is a scalar, (4.4b), 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 (4.4) as

$$\widetilde{\tau}_i^{-1}(1 + 2\gamma\tau_i)\tau_i^{-1} - \widetilde{\tau}_{i+1}\tau_{i+1}^{-1} \ge 0, \tag{4.9a}$$

$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1} \ge -2\widetilde{\tau}_i^{-1}\gamma_i^\perp, \tag{4.9b}$$

$$\frac{1}{1 - \delta}\left(\eta_i\|KP\|^2 + (\widetilde{\tau}_i^{-1}\tau_i^\perp - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^\perp)\|K\|^2\right) \ge -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \tag{4.9c}$$

We are almost ready to state a general convergence result for projective Γ. However, we want to make one more thing more explicit. Since $\Gamma_i - \Gamma = \gamma_i^\perp P^\perp$ and $R_{i+1} = \rho_{i+1}I$, we suppose for simplicity that

$$\phi_{R_{i+1}}(y) = \rho_{i+1}\phi(y) \quad \text{and} \quad \psi_{\Gamma_i - \Gamma}(x) = \gamma_i^\perp\psi^\perp(P^\perp x) \tag{4.10}$$

for some $\phi : Y \to \mathbb{R}$ and $\psi^\perp : P^\perp X \to \mathbb{R}$. 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

$$G(x') - G(x) \geq \langle \partial G(x), x' - x \rangle + \frac{\bar{\gamma}}{2} \|P(x' - x)\|^2 + \frac{\gamma^\perp}{2} \left(\|P^\perp(x' - x)\|^2 - \psi(P^\perp(x' - x)) \right),$$
$$(x, x' \in X; \ 0 \leq \gamma^\perp \leq \bar{\gamma}^\perp), \quad \text{(G-pcr)}$$

and

$$F^*(y') - F^*(y) \geq \langle \partial F^*(y), y' - y \rangle + \frac{\rho}{2} \left(\|y' - y\|^2 - \phi(y' - y) \right),$$
$$(y, y' \in Y; \ 0 \leq \rho \leq \bar{\rho}). \quad \text{(F*-pcr)}$$

Analogous variants 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* (4.9) *is satisfied with*

$$0 \leq \gamma_i^\perp \leq \lambda \bar{\gamma}^\perp, \quad 0 \leq \rho_i \leq \lambda \bar{\rho}, \quad \text{and} \quad \widetilde{\tau}_0 \geq \widetilde{\tau}_i > 0. \tag{4.11}$$

If we solve (4.9a) *exactly, define* T_i, Γ_i, *and* Σ_{i+1} *through* (4.6) *and* (4.8), *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* (2.13) *the estimate*

$$\frac{\delta}{2} \|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + 2\gamma} \mathcal{G}^N \leq \widetilde{\tau}_N \tau_N \left(C_0 + \sum_{i=0}^{N-1} D_{i+1} \right). \tag{4.12}$$

If we take $\lambda = 1$, *then* (4.12) *holds with* $\mathcal{G}^N = 0$.

Observe that presently

$$D_{i+1} = \widetilde{\tau}_i^{-1} \gamma_i^\perp \psi(x^{i+1} - \widehat{x}) + \widetilde{\tau}_{i+1}^{-1} \rho_{i+1} \phi(y^{i+1} - \widehat{y}). \tag{4.13}$$

Proof. As we have assumed through (4.11), or otherwise already verified its conditions, we may apply Theorem 3.1. Multiplying (3.4) by $\widetilde{\tau}_N \tau_N$, we obtain

$$\frac{\delta}{2} \|x^N - \widehat{x}\|_P^2 + \widetilde{q}_N \widetilde{\tau}_N \tau_N \mathcal{G}^N \leq \widetilde{\tau}_N \tau_N \left(C_0 + \sum_{i=0}^{N-1} D_{i+1} \right). \tag{4.14}$$

Now, observe that solving (4.9a) exactly gives

$$\widetilde{\tau}_N^{-1} \tau_N^{-1} = \widetilde{\tau}_{N-1}^{-1} \tau_{N-1}^{-1} + 2\gamma \widetilde{\tau}_{N-1}^{-1} = \widetilde{\tau}_0^{-1} \tau_0^{-1} + \sum_{j=0}^{N-1} 2\gamma \widetilde{\tau}_j^{-1} = \widetilde{\tau}_0^{-1} \tau_0^{-1} + 2\gamma \widetilde{q}_N. \tag{4.15}$$

Therefore, we have the estimate

$$\widetilde{q}_N \widetilde{\tau}_N \tau_N = \frac{\widetilde{q}_N}{\widetilde{\tau}_0^{-1} \tau_0^{-1} + 2\gamma \widetilde{q}_N} = \frac{1}{\widetilde{\tau}_0^{-1} \tau_0^{-1} \widetilde{q}_N^{-1} + 2\gamma} \geq \frac{1}{\tau_0^{-1} + 2\gamma}. \tag{4.16}$$

With this, (4.14) yields (4.12). $\qquad\square$

17

4.3   Primal and dual penalties with projective Γ

The right-hand-sides of (4.9b) and (4.9c) relate to D_{i+1} in the convergence estimate (4.12). 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} in (3.4). Optimal rates are therefore obtained by solving (4.9b) and (4.9c) 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 (4.9c) by the simpler condition

$$\frac{1}{1-\delta}(\widetilde{\tau}_i^{-1}\tau_i^{\perp} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp})\|K\|^2 \geq -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \tag{4.17}$$

Then we try to make the left hand sides of (4.9b) and (4.17) proportional with only $\tau_{i+1}^{\perp}$ as a free variable. That is, for some proportionality constant $\zeta > 0$, we solve

$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1} = \zeta(\widetilde{\tau}_i^{-1}\tau_i^{\perp} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp}). \tag{4.18}$$

Multiplying both sides of (4.18) 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

$$\tau_{i+1}^{\perp} = \frac{1}{2}\left(\widetilde{\omega}_i(\tau_i^{\perp} - \zeta^{-1}\tau_i^{\perp,-1}) + \sqrt{\widetilde{\omega}_i^2(\tau_i^{\perp} - \zeta^{-1}\tau_i^{\perp,-1})^2 + 4\zeta^{-1}}\right). \tag{4.19}$$

Solving (4.9b) and (4.17) as equalities, (4.18) and (4.19) give

$$2\widetilde{\tau}_i^{-1}\gamma_i^{\perp} = \frac{2\zeta(1-\delta)}{\|K\|^2}\widetilde{\tau}_{i+1}^{-1}\rho_{i+1} = \zeta(\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp} - \widetilde{\tau}_i^{-1}\tau_i^{\perp}). \tag{4.20}$$

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

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

Thus $\omega_i^{\perp} \geq \widetilde{\omega}_i$ if

$$(1 - \zeta^{-1}\tau_i^{\perp,-2})^2 + 4\zeta^{-1}\widetilde{\omega}_i^{-2}\tau_i^{\perp,-2} \geq (1 + \zeta^{-1}\tau_i^{\perp,-2})^2.$$

This gives the condition $\zeta^{-1}\widetilde{\omega}_i^{-2}\tau_i^{\perp,-2} \geq \zeta^{-1}\tau_i^{\perp,-2}$, which says that (4.20) 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 (4.19), and set*

$$\widetilde{\omega}_i = \omega_i = 1/\sqrt{1 + 2\gamma\tau_i}. \tag{4.21}$$

Then $\omega_i^{\perp} \geq \widetilde{\omega}_i$, $\widetilde{\tau}_i \leq \widetilde{\tau}_0$, and (4.9) is satisfied with the right-hand-sides given by the non-negative quantity in (4.20). Moreover,

$$\tau_i^{\perp} \leq \zeta^{-1/2} \implies \tau_{i+1}^{\perp} \leq \zeta^{-1/2}. \tag{4.22}$$

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 (4.25).

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

$$\omega_i = 1/\sqrt{1 + 2\gamma\tau_i}, \quad \text{and}$$

$$\omega_i^\perp = \frac{1}{2}\left((1 - \zeta^{-1}\tau_i^{\perp, -2})\omega_i + \sqrt{(1 - \zeta^{-1}\tau_i^{\perp, -2})^2\omega_i^2 + 4\zeta^{-1}\tau_i^{\perp, -2}}\right).$$

4:    Update

$$\tau_{i+1} = \tau_i\omega_i, \quad \tau_{i+1}^\perp = \tau_i^\perp\omega_i^\perp, \quad \text{and}$$

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

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

$$x^{i+1} := (I + T_i\partial G)^{-1}(x^i - T_i K^* y^i), \qquad (4.23a)$$

$$\bar{x}^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \qquad (4.23b)$$

$$y^{i+1} := (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1}K\bar{x}^{i+1}). \qquad (4.23c)$$

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

Proof. The choice (4.21) satisfies (4.9a), so that (4.9) in its entirety will be satisfied with the right-hand sides of (4.9b)–(4.9c) given by (4.20). The bound $\widetilde{\tau}_i \leq \widetilde{\tau}_0$ follows from $\widetilde{\omega}_i \leq 1$. Finally, the implication (4.22) is a simple estimation of (4.19). □

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 (4.20).

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 (4.19) and (4.21) that

$$
\begin{aligned}
\gamma_i^{\perp} = \frac{\zeta}{2}(\widetilde{\omega}_i^{-1}\tau_{i+1}^{\perp} - \tau_i^{\perp}) &= \frac{1}{2}\left(-\zeta\tau_i^{\perp} - \tau_i^{\perp,-1} + \sqrt{(\zeta\tau_i^{\perp} - \tau_i^{\perp,-1})^2 + 4\zeta\widetilde{\omega}_i^{-2}}\right) \\
&\leq \frac{1}{2}\sqrt{(\zeta\tau_i^{\perp} - \tau_i^{\perp,-1})^2 - (\zeta\tau_i^{\perp} + \tau_i^{\perp,-1})^2 + 4\zeta\widetilde{\omega}_i^{-2}} \\
&= \sqrt{\zeta(\widetilde{\omega}_i^{-2} - 1)} = \sqrt{2\zeta\gamma\tau_i} \leq \sqrt{2\zeta\gamma\tau_0}.
\end{aligned}
\tag{4.24}
$$

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 (4.20), (4.21), and (4.24) that

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

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 get on τ_0 and ζ the always satisfiable condition

$$
0 < \tau_0 \leq \frac{\lambda^2}{2\gamma}\min\left\{\frac{\bar{\gamma}^{\perp,2}}{\zeta}, \frac{\bar{\rho}^2\zeta(1-\delta)^2}{\|K\|^4}\right\}.
\tag{4.25}
$$

If now $\phi \equiv C_\phi$ and $\psi \equiv C_\psi^{\perp}$, using the non-negativity of (4.20) when $0 < \widetilde{\omega}_i \leq 1$, we may calculate

$$
\sum_{i=0}^{N-1}\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}\phi(y^{i+1} - \widehat{y}) = \frac{\|K\|^2C_\phi}{2(1-\delta)}\left(\sum_{i=0}^{N-1}\frac{\widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp}}{2} - \sum_{i=0}^{N-1}\frac{\widetilde{\tau}_i^{-1}\tau_i^{\perp}}{2}\right) \leq \frac{\|K\|^2C_\phi}{2(1-\delta)}\widetilde{\tau}_N^{-1}\tau_N^{\perp}.
\tag{4.26}
$$

Similarly

$$
\sum_{i=0}^{N-1}\widetilde{\tau}_i^{-1}\gamma_i^{\perp}\psi(x^{i+1} - \widehat{x}) \leq \frac{\zeta C_\psi^{\perp}}{2}\widetilde{\tau}_N^{-1}\tau_N^{\perp}.
\tag{4.27}
$$

Using these expression to expand (4.13), 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 $\bar{\gamma}, \bar{\gamma}^{\perp}, \bar{\rho} > 0$, and*

$$
\phi \equiv C_\phi, \quad \text{and} \quad \psi \equiv C_\psi^{\perp}
$$

for some constants $C_\phi, C_\psi^{\perp} > 0$. With $\lambda = 1/2$, fix $\gamma \in (0, \lambda\gamma]$. Select initial $\tau_0, \tau_0^{\perp} > 0$, as well as $\delta \in (0,1)$ and $\zeta \leq (\tau_0^{\perp})^{-2}$ satisfying (4.25). Then Algorithm 3 satisfies for some $C_0, C_\tau > 0$ the estimate

$$
\frac{\delta}{2}\|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + 2\gamma}\mathcal{G}^N \leq \frac{C_0C_\tau^2}{N^2} + \frac{C_\tau}{2N}\left(\zeta^{1/2}C_\psi^{\perp} + \frac{\zeta^{-1/2}\|K\|^2}{1-\delta}C_\phi\right), \quad (N \geq 0).
\tag{4.28}
$$

If we take $\lambda = 1$, then (4.12) holds with $\mathcal{G}^N = 0$.

Proof. During the course of the derivation of Algorithm 3, we have verified (4.9), solving (4.9a) as an equality. Moreover, Lemma 4.1 and (4.25) guarantee (4.11). We may therefore apply Proposition 4.1. Inserting (4.26) and (4.27) into (4.12) and (4.13) gives

$$\frac{\delta}{2}\|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + 2\gamma}\mathcal{G}^N \leq \tau_N\widetilde{\tau}_N\left(C_0 + \frac{\zeta C_\psi^\perp}{2}\widetilde{\tau}_N^{-1}\tau_N^\perp + \frac{\|K\|^2 C_\phi}{2(1-\delta)}\widetilde{\tau}_N^{-1}\tau_N^\perp\right). \quad (4.29)$$

The condition $\zeta \leq (\tau_0^\perp)^{-2}$ now guarantees $\tau_N^\perp \leq \zeta^{-1/2}$ through (4.22). 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, (4.28) follows by simple estimation of (4.29). □

Remark 4.2. As a special case of Algorithm 3, if we choose $\zeta = \tau_0^{\perp,-2}$, then we can show from (4.19) that $\tau_i^\perp = \tau_0^\perp = \zeta^{-1/2}$ for all $i \in \mathbb{N}$.

4.4 Dual penalty only with projective Γ

Continuing with the projective Γ setup of Section 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 (4.11) and (4.9), with (4.9a) exactly. Since $\gamma_i^\perp = 0$, (4.9b) becomes

$$\widetilde{\tau}_i^{-1}\tau_i^{\perp,-1} - \widetilde{\tau}_{i+1}^{-1}\tau_{i+1}^{\perp,-1} \geq 0. \quad (4.30)$$

With respect to $\tau_{i+1}^\perp$, the left hand side of (4.9c) is maximised (and the penalty on the right hand side minimised) when (4.30) is minimised. Thus we solve (4.30) exactly, which gives

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

In consequence $\omega_i^\perp = \widetilde{\omega}_i^{-1}$, and (4.9c) becomes

$$\frac{1}{1-\delta}\eta_i\|KP\|^2 + \frac{\widetilde{\tau}_i^{-2}}{1-\delta}(1 - \widetilde{\omega}_i^{-2})\|K\|^2 \geq -2\widetilde{\tau}_{i+1}^{-1}\rho_{i+1}. \quad (4.31)$$

Here $\eta_i \geq 0$, so we estimate this as $\eta_i = 0$ as in (4.17). This suggests to choose

$$\widetilde{\omega}_i := \frac{1}{1 + a_i\widetilde{\tau}_i^2} \quad \text{and} \quad \omega_i := \frac{1}{\widetilde{\omega}_i(1 + 2\gamma\tau_i)}, \quad (4.32)$$

for some, yet undetermined, $a_i > 0$. Solving (4.31) as an equality for ρ_{i+1}, then

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

This needs $\rho_{i+1} \leq \lambda\bar{\rho}$. Since $\widetilde{\tau}_i^{-1} \geq \widetilde{\tau}_0^{-1}$, we can satisfy this for large enough i if $a_i \searrow 0$, or generally if $\widetilde{\tau}_0$ is small enough and $\{a_i\}$ non-decreasing. In particular, if $\{a_i\}$ is descending, it suffices

$$a_0\tau_0^\perp\widetilde{\tau}_0^2\frac{\|K\|^2}{2(1-\delta)} \leq \bar{\rho} \quad (4.33)$$

Noting that (4.32) ensures $\widetilde{\tau}_{i+1}^{-2} = \widetilde{\tau}_i^{-2} + a_i$, we see that

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

Assuming ϕ to have the structure (4.10), 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 (4.12) in Proposition 4.1 states

$$\frac{\delta}{2}\|P(x^N - \widehat{x})\|^2 + \frac{1}{\tau_0^{-1} + 2\gamma} \mathcal{G}^N \leq \mu_0^N C_0 + \frac{\|K\|^2}{2(1-\delta)} \mu_1^N \tag{4.34}$$

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 (4.34) 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 $q > 0$ and

$$a_i := \widetilde{\tau}_0^{-2}\left((i+1)^q - i^q\right). \tag{4.35}$$

This gives

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

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

$$\mu_0^N \leq \frac{\widetilde{\tau}_0 C_a}{N^{1+q/2}}, \quad \text{and} \quad \mu_1^N \leq \frac{C_a C_\phi}{\widetilde{\tau}_0 N^{1-q/2}}. \tag{4.36}$$

The choice $q = 0$ gives uniform $O(1/N)$ over both the initialisation and the dual sequence. By choosing $q < 2$ large, we can get arbitrarily close to $O(1/N^2)$ rate with respect to the initialisation, at the cost of the rate μ_1^N with respect to the dual sequence becoming closer and closer to zero.

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 decreasing sequence $\{a_i\}_{i=0}^{\infty}$, for example as in (4.35). Initial step parameters $\tau_0, \tau_0^{\perp}, \widetilde\tau_0 > 0$, as well as $\delta \in (0, 1)$, satisfying (4.33).

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

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

   as well as

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

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

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

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* (4.35) *for some* $q > 0$. *Select initial* $\tau_0, \tau_0^{\perp}, \widetilde\tau_0 > 0$ *and* $\delta \in (0, 1)$ *verifying* (4.33). *Then Algorithm 4 satisfies*

$$\frac{\delta}{2}\|P(x^N - \widehat x)\|^2 + \frac{1}{\tau_0^{-1} + \gamma}\mathcal{G}^N \leq \frac{\widetilde\tau_0 C_a C_0}{N^{1+q/2}} + \frac{C_a C_{\phi}\|K\|^2}{2(1 - \delta)\widetilde\tau_0^2 N^{1-q/2}}, \quad (N \geq 2). \quad (4.38)$$

If we take $\lambda = 1$, *then* (4.38) *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 (4.32) that forces $\widetilde\omega_i \leq 1$. Also, have already derived the rate (4.34) from (4.12). Inserting (4.36) into (4.34), noting that the former is only valid for $N \geq 2$, immediately gives (4.38) $\qquad\square$

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
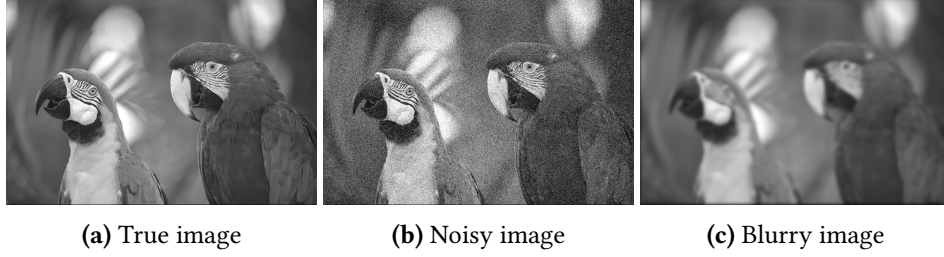
(a) True image   (b) Noisy image   (c) Blurry image

Figure 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/.

work form Hilbert spaces to Banach spaces, as was done in [20] for the original method of [11]. 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

$$\min_{x \in X} \ G_0(f - Ax) + \alpha F(Kx). \tag{5.1}$$

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, although in case of support vector machines, for example, the smooth function is the 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 our algorithm can provide improved convergence rates.

In image processing examples abound; we refer to [12] for an overview. In total variation (TV) regularisation we still take $F(x) = \|x\|_1$, but $K = \nabla$. Strictly speaking, this has to be formulated in the Banach space BV(Ω), 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 sub-sampling 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, e.g., [6, 7]), where we take $A = S\mathcal{F}$ for $\mathcal{F}$ 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 $A = \mathcal{F}^* \hat{a} \mathcal{F}$, multiplying the Fourier transform by some $\hat{a} \in L^2(\Omega)$. If $|\hat{a}| \geq \gamma$ on a sub-domain, 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 (5.1). Thus, for some $M > 0$, we can add to (5.1) the artificial constraint

$$G'(x) := \delta_{\|\cdot\|_Z \leq M}(P^\perp x). \tag{5.2}$$

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 Poincaré's inequality [1] on $BV(\Omega)$ 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 (5.2) with $Z = L^2$ to (5.1).

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 (5.2) is never active. Indeed, A having a non-trivial nullspace also causes duality gaps for (P) to be numerically infinite. In [32] 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.

5.2   TGV^2 denoising and related problem structure

So far, we have considered very simple regularisation terms. Total generalised variation, TGV, was introduced in [8] 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 BV(Ω). For two parameters $\alpha, \beta > 0$, the regularisation functional is written in the differentiation cascade form of [9] 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 $x = (u, w)$ and $y = (y_1, y_2)$, we may write the problem

$$\min_u G_0(f - Au) + \mathrm{TGV}^2_{(\beta, \alpha)}(u), \tag{5.3}$$

in the saddle-point form (P) with

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

If $A = I$, as is the case for denoising, this is an instance of the general structure

$$G(x_1, x_2) = G_1(x_1) + G_2(x_2), \quad F^*(y_1, y_2) = F_1^*(y_1) + F_2^*(y_2), \quad \text{and} \quad K := \begin{pmatrix} K_{1,1} & K_{1,2} \\ 0 & K_{2,2} \end{pmatrix},$$

where G_1 is strongly convex with factor γ. To apply Algorithm 3, we therefore need to find C_ψ and $\bar{\gamma}^\perp$ satisfying for all $0 \le \gamma^\perp \le \bar{\gamma}^\perp$ the condition

$$G_2(x_2') - G_2(x_2) \ge \langle \partial G_2(x_2), x_2' - x_2 \rangle + \frac{\gamma^\perp}{2}(\|x_2' - x_2\|^2 - C_\psi), \quad (x_2', x_2 \in X_2). \tag{5.4}$$

For both Algorithm 4 and Algorithm 3, we also need F_j^*, $(j = 1, 2)$, to satisfy for all $0 \le \rho \le \bar{\rho}$ and some C_ϕ the condition

$$F_j^*(y_j') - F_j^*(y_j) \ge \langle \partial F_j^*(y_j), y_j' - y_j \rangle + \frac{\rho}{2}(\|y_j' - y_j\|^2 - C_\phi), \quad (y_j', y_j \in Y_j). \tag{5.5}$$

If these conditions hold, we have

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

As this is compatible with the splitting of G into G_1 and G_2, the prox-update (4.23a) splits into the uncoupled updates

$$x_1^{i+1} = (I + \tau_i G_1^*)^{-1}(x_1^i - \tau_i K_{1,1}^* y_1^i),$$
$$x_2^{i+1} = (I + \tau_i^\perp G_2^*)^{-1}(x_2^i - \tau_i^\perp K_{1,2}^* y_1^i - \tau_i^\perp K_{2,2}^* y_2^i).$$

For the general class of problems, (F*-pcr) with $\bar{\rho} = \infty$ is immediate from Lemma 2.1. For TGV^2 denoising in particular, the Sobolev–Korn inequality [29] allows us to bound on a connected domain $\Omega \subset \mathbb{R}^2$ an optimal $\hat{w}$ to (5.3) 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 (5.3). Therefore we may again add the artificial constraint $G_2(w) = \delta_{\|\cdot\|_{L^2} \le M}(w)$ to the TGV^2 denoising problem. By Lemma 2.1, G will then satisfy (G-pcr) with $\bar{\gamma}^\perp = \infty$.

(a) τ **(b)** $\tau^\perp$

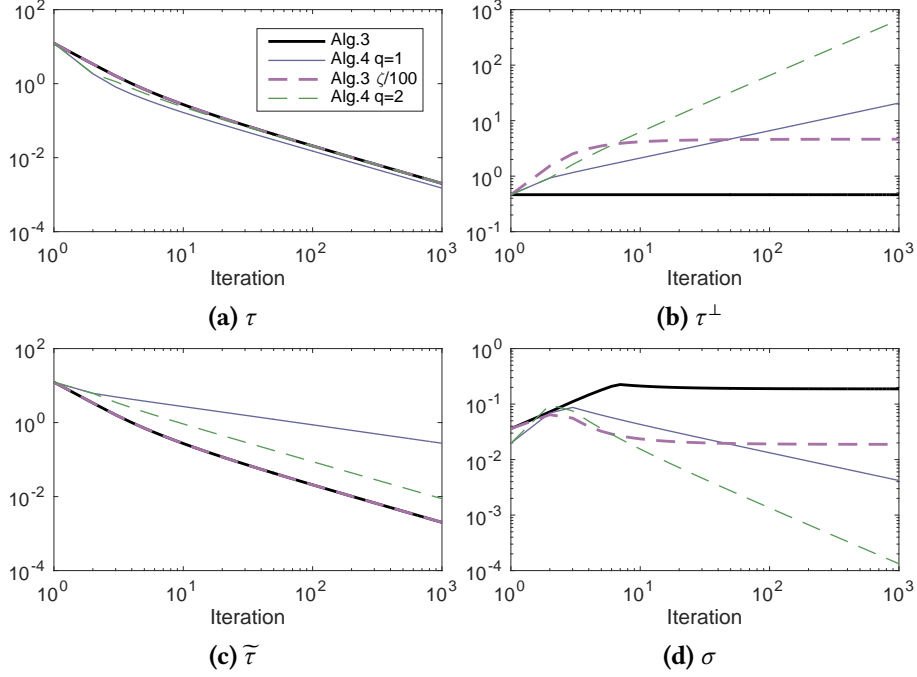**(c)** $\widetilde{\tau}$ **(d)** σ

Figure 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=2' uses the value $q = 2$ for Algorithm 4, everything else being kept equal.

5.3 Numerical results

We demonstrate our algorithms on TGV^2 denoising and TV deblurring. Our tests are done on the photographs in Figure 1, both at the original resolution of 768×512, and scaled down by a factor of 0.25 to 192×128 pixels. For both of our example problems, we calculate a target solution by taking one million iterations of the basic PDHGM (2.1). We also tried interior point methods for this, but they are only practical for the smaller denoising problem.

We evaluate Algorithm 3 and 4 against the standard unaccelerated PDHGM of [11], as well as the mixed-rate method of [14], denoted here C-L-O. 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 [32], because the real duality gap is in practise infinite when A has a non-trivial nullspace. We do this dynamically, upgrading the M in (5.2) every time we compute the duality gap. For both of our example problems, we use for simplicity $Z = L^2$ in (5.2). In the calculation of the final duality gaps comparing each
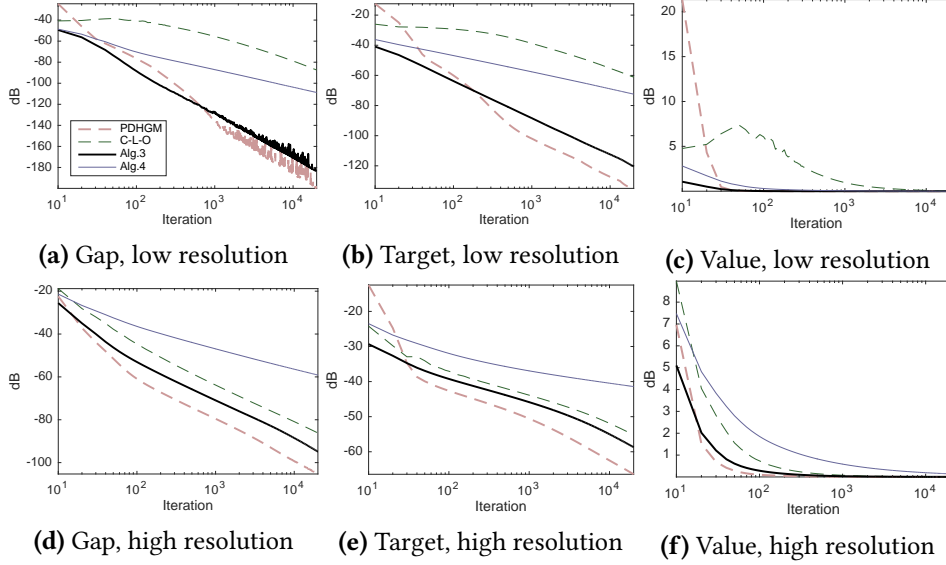
(a) Gap, low resolution   **(b)** Target, low resolution   **(c)** Value, low resolution

(d) Gap, high resolution   **(e)** Target, high resolution   **(f)** Value, high resolution

Figure 3: TGV^2 denoising performance, 20000 iterations, high and low resolution images. The plot is logarithmic, with the decibels calculated as in Section 5.3.

	low resolution						high resolution					
	gap $\leq$ −50dB		tgt $\leq$ −50dB		val $\leq$ 1dB		gap $\leq$ −50dB		tgt $\leq$ −50dB		val $\leq$ 1dB	
Method	iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
PDHGM	30	0.37s	50	0.50s	30	0.37s	50	7.30s	870	125.87s	30	4.29s
C-L-O	500	4.67s	5170	51.78s	970	9.04s	190	37.47s	6400	1261.36s	80	15.76s
Alg.3	20	0.29s	30	0.36s	20	0.29s	80	12.30s	3320	512.35s	40	6.20s
Alg.4	20	0.40s	200	1.92s	40	0.62s	2080	317.93s	−	−	340	52.06s

Table 1: TGV^2 denoising performance, maximum 20000 iterations. 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.

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 12dB. In the downscaled image, these become, respectively, 6.15 or 25.7dB. 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
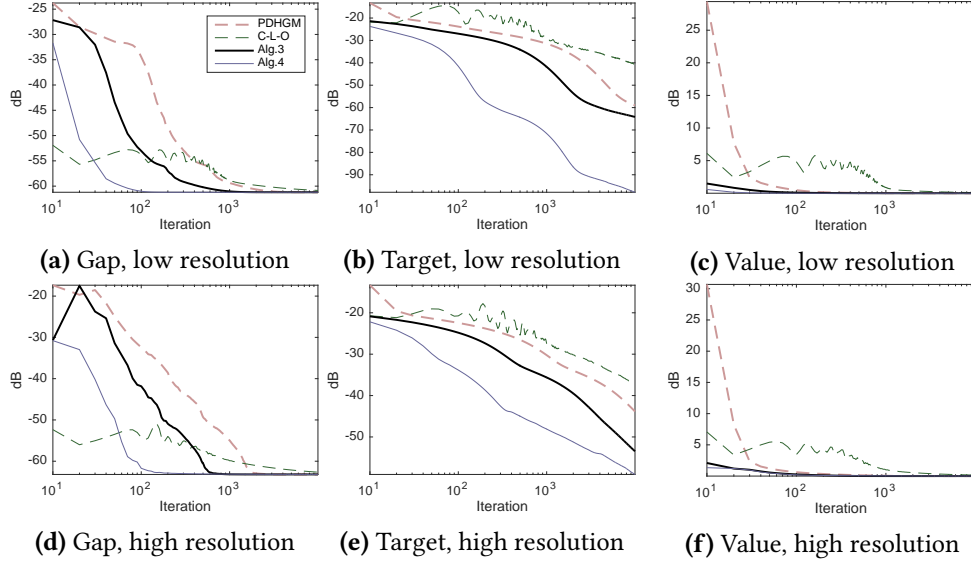
(a) Gap, low resolution **(b)** Target, low resolution **(c)** Value, low resolution

(d) Gap, high resolution **(e)** Target, high resolution **(f)** Value, high resolution

Figure 4: TV deblurring performance, 10000 iterations, high and low resolution images. The plot is logarithmic, with the decibels calculated as in Section 5.3.

0.25 downscaling factor. See [15] 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 $\|K\|^2 \leq 11.4$ [32]. 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 $q = 1$ 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 Figure 2. For the C-L-O, we use the detailed parametrisation from [13, 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.

We take fixed 20000 iterations, and initialise each algorithm with $y^0 = 0$ and $x^0 = (u^0, w^0) = (f, 0)$. 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 Figure 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 PDHGM starts to overtake. This

29

	low resolution						high resolution					
	gap ≤ -50dB		tgt ≤ -50dB		val ≤ 1dB		gap ≤ -50dB		tgt ≤ -50dB		val ≤ 1dB	
Method	iter	time	iter	time	iter	time	iter	time	iter	time	iter	time
PDHGM	230	1.40s	4800	31.16s	60	0.41s	540	46.09s	–	–	70	6.04s
C-L-O	10	0.12s	–	–	950	5.99s	10	1.07s	–	–	1000	94.49s
Alg.3	80	0.64s	1630	13.56s	20	0.21s	180	21.41s	6760	799.94s	40	4.89s
Alg.4	20	0.21s	140	1.17s	10	0.14s	60	7.30s	1550	190.39s	30	3.69s

Table 2: TV deblurring performance, maximum 10000 iterations. 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.

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 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 to the other algorithms. Overall, Algorithm 3 gives good results fast, although the basic PDHGM seems to perform, in practise, better asymptotically.

TV *deblurring*   Our test image has now been distorted by Gaussian blur of kernel width 4, which we intent to remove. We denote by $\hat{a}$ the Fourier presentation of the blur operator as discussed in Section 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 $P = \mathcal{F}^*\hat{p}\mathcal{F}$.

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 is exactly as TGV^2 denoising above, of course with appropriate Ω_U and $\|K\|^2 \leq 8$ for $K = \nabla$ [10]. As initial iterates we use $x^0 = f$ the blurry image, and $y^0 = 0$.

The results are Figure 4 and Table 2. It does not appear numerically feasible to go significantly below -60dB 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. We note excellent performance from Algorithm 3, and this time from Algorithm 4 as well.

Acknowledgements

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 Kodak image suite was in particular available at `http://r0k.us/graphics/kodak/`. Our implementations of the proposed algorithms will be archived as the final version of the paper is submitted.

References

[1] L. Ambrosio, N. Fusco and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, 2000.

[2] A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Transactions on Image Processing **18** (2009), 2419–2434, doi:10.1109/TIP.2009.2028250.

[3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences **2** (2009), 183–202, doi:10.1137/080716542.

[4] A. Beck and M. Teboulle, *Smoothing and first order methods: A unified framework*, SIAM Journal on Optimization **22** (2012), 557–580, doi:10.1137/100818327.

[5] A. Beck and M. Teboulle, *A fast dual proximal gradient algorithm for convex minimization and applications*, Operations Research Letters **42** (2014), 1 – 6, doi:10.1016/j.orl.2013.10.007.

[6] M. Benning, L. Gladden, D. Holland, C.-B. Schönlieb and T. Valkonen, *Phase reconstruction from velocity-encoded MRI measurements – A survey of sparsity-promoting variational approaches*, Journal of Magnetic Resonance **238** (2014), 26–43, doi:10.1016/j.jmr.2013.10.003.
URL `http://iki.fi/tuomov/mathematics/phaserec.pdf`

[7] M. Benning, F. Knoll, C.-B. Schönlieb and T. Valkonen, *Preconditioned ADMM with nonlinear operator constraint* (2015), submitted, arXiv:1511.00425.
URL `http://iki.fi/tuomov/mathematics/nonlinearADMM.pdf`

[8] K. Bredies, K. Kunisch and T. Pock, *Total generalized variation*, SIAM Journal on Imaging Sciences **3** (2011), 492–526, doi:10.1137/090769521.

[9] K. Bredies and T. Valkonen, *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.
URL http://iki.fi/tuomov/mathematics/SampTA2011.pdf

[10] A. Chambolle, *An algorithm for mean curvature motion*, Interfaces and Free Boundaries **6** (2004), 195.

[11] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision **40** (2011), 120–145, doi:10.1007/s10851-010-0251-1.

[12] T. Chan and J. Shen, *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*, Society for Industrial and Applied Mathematics (SIAM), 2005.

[13] K. Chen and D. A. Lorenz, *Image sequence interpolation using optimal control*, Journal of Mathematical Imaging and Vision **41** (2011), 222–238, doi:10.1007/s10851-011-0274-2.

[14] Y. Chen, G. Lan and Y. Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, SIAM Journal on Optimization **24** (2014), 1779–1814, doi:10.1137/130919362.

[15] J. C. de Los Reyes, C.-B. Schönlieb and T. Valkonen, *Bilevel parameter learning for higher-order total variation regularisation models* (2015), submitted, arXiv:1508.07243.
URL http://iki.fi/tuomov/mathematics/tgv_learn.pdf

[16] I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1999.

[17] E. Esser, X. Zhang and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM Journal on Imaging Sciences **3** (2010), 1015–1046, doi:10.1137/09076934X.

[18] O. Fercoq and P. Bianchi, *A coordinate descent primal-dual algorithm with large step size and possibly non separable functions* (2015), arXiv:1508.04625.

[19] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective*, SIAM Journal on Imaging Sciences **5** (2012), 119–149, doi:10.1137/100814494.

[20] T. Hohage and C. Homann, *A generalization of the Chambolle-Pock algorithm to Banach spaces with applications to inverse problems* (2014), preprint, arXiv:1412.0126.

[21] D. Lorenz and T. Pock, *An inertial forward-backward algorithm for monotone inclusions*, Journal of Mathematical Imaging and Vision **51** (2015), 311–325, doi:10.1007/s10851-014-0523-2.

[22] T. Möllenhoff, E. Strekalovskiy, M. Moeller and D. Cremers, *The primal-dual hybrid gradient method for semiconvex splittings*, SIAM Journal on Imaging Sciences **8** (2015), 827–857, doi:10.1137/140976601.

[23] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Mathematics Doklady **27** (1983), 372–376.

[24] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming **103** (2005), 127–152, doi:10.1007/s10107-004-0552-5.

[25] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, in: *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, 1762 –1769, doi:10.1109/ICCV.2011.6126441.

[26] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, 1998.

[27] W. Rudin, *Functional Analysis*, International series in pure and applied mathematics, McGraw−Hill, 2006.

[28] S. Setzer, *Operator splittings, Bregman methods and frame shrinkage in image processing*, International Journal of Computer Vision **92** (2011), 265–280, doi:10.1007/s11263-010-0357-3.

[29] R. Temam, *Mathematical problems in plasticity*, Gauthier-Villars, 1985.

[30] T. Valkonen, *A primal-dual hybrid gradient method for non-linear operators with applications to MRI*, Inverse Problems **30** (2014), 055012, doi:10.1088/0266-5611/30/5/055012, arXiv:1309.5032.
URL http://iki.fi/tuomov/mathematics/nl-pdhgm.pdf

[31] T. Valkonen, *Optimising big images*, in: *Big Data Optimization: Recent Developments and Challenges*, Edited by A. Emrouznejad, Studies in Big Data, Springer-Verlag, 2015, accepted.

[32] T. Valkonen, K. Bredies and F. Knoll, *Total generalised variation in diffusion tensor imaging*, SIAM Journal on Imaging Sciences **6** (2013), 487–525, doi:10.1137/120867172.
URL http://iki.fi/tuomov/mathematics/dtireg.pdf