INERTIAL, CORRECTED, PRIMAL-DUAL PROXIMAL SPLITTING

Tuomo Valkonen*

2018-04-24 (revised 2020-02-12)

Abstract We study inertial versions of primal-dual proximal splitting, also known as the Chambolle–Pock method. Our starting point is the preconditioned proximal point formulation of this method. By adding correctors corresponding to the anti-symmetric part of the relevant monotone operator, using a FISTA-style gap unrolling argument, we are able to derive gap estimates instead of merely ergodic gap estimates. Moreover, based on adding a diagonal component to this corrector, we are able to combine strong convexity based acceleration with inertial acceleration. We test our proposed method on image processing and inverse problems problems, obtaining convergence improvements for sparse Fourier inversion and Positron Emission Tomography.

1 INTRODUCTION

For convex, proper, and lower semicontinuous $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$, and a bounded linear operator $K \in \mathcal{L}(X; Y)$ on Hilbert spaces X and Y, we will derive inertial primal-dual optimisation methods for the problem

$$\min_{x \in X} G(x) + F(Kx). \tag{1.1}$$

If K is the identity, and F is smooth, a classical algorithm for the iterative solution of (1.1) is the forward–backward splitting method $x^{i+1} := \operatorname{prox}_{\tau G}(x^i - \tau \nabla F(x^i))$, where $\tau L < 1$ for L the Lipschitz factor of ∇F. That is, we take proximal steps with respect to G, and gradient steps with respect to F. The proximal step needs to be efficiently realisable, i.e., G needs to be "prox-simple". If no strong convexity is present, the iterates of the forward–backward splitting generally converge weakly, and the function values at the rate $O(1/N)$. By applying *inertia*, the latter can be improved to $O(1/N^2)$. This in essence consists of rebasing the algorithm at an inertial variable $\bar{x}^i$:

$$x^{i+1} := \operatorname{prox}_{\tau G}(\bar{x}^i - \tau \nabla F(\bar{x}^i)), \quad \text{where} \quad \bar{x}^i := (1 + \alpha_i)x^i - \alpha_i x^{i-1} \tag{1.2}$$

*Department of Mathematics and Statistics, University of Helsinki, Finland *and* ModeMat, Escuela Politécnica Nacional, Quito, Ecuador; *previously* Department of Mathematical Sciences, University of Liverpool, United Kingdom. tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

for suitable inertial parameters $\{\alpha_i\}_{i\in\mathbb{N}}$. In FISTA [3], which itself is an extension of Nesterov's accelerated gradient descent [20], one would take

$$\alpha_{i+1} := \lambda_{i+1}(\lambda_i^{-1} - 1) \quad \text{for} \quad \lambda_{i+1}^{-1} := \sqrt{\lambda_i^{-2} + 1/4} + 1/2. \tag{1.3}$$

For this scheme, no convergence rates of the iterates themselves are known, although weak convergence can be obtained with small modifications [8]. Several studies have sought to further optimise the inertial parameters; we refer merely to a few of the most recent works [2, 16] and references therein.

If F is non-smooth, but G is smooth, we can apply forward–backward splitting or FISTA with the roles of the two functions exchanged. However, if K is not the identity, $F \circ K$ is rarely prox-simple, so these methods are not practically applicable. Nevertheless, denoting by F^* the Fenchel conjugate of F, we can reformulate (1.1) as

$$\min_{x\in X} \max_{y\in Y} \; G(x) + \langle Kx, y\rangle - F^*(y). \tag{1.4}$$

A popular iterative method for this class of problems is the *primal-dual proximal splitting* (PDPS), commonly known as the Chambolle–Pock method [7]. It takes alternate proximal steps with respect to the primal and dual variables x and y:

$$\begin{cases} x^{i+1} := \operatorname{prox}_{\tau_i G}(x^i - \tau_i K^* y^i), \\ x_\omega^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \\ y^{i+1} := \operatorname{prox}_{\sigma_{i+1}F^*}(y^i + \sigma_{i+1} K x_\omega^{i+1}). \end{cases} \tag{1.5}$$

In the basic version the over-relaxation parameter $\omega_i \equiv 1$, and the primal and dual step lengths $\tau_i \equiv \tau_0$, $\sigma_i \equiv \sigma_0$ with $\tau_0\sigma_0\|K\|^2 < 1$. This yields $O(1/N)$ convergence rate for an ergodic gap functional, and weak convergence of the iterates. If G is strongly convex with factor $\gamma > 0$, an accelerated version updates $\tau_{i+1} := \omega_i\tau_i$ and $\sigma_{i+1} := \sigma_i/\omega_i$ for $\omega_i := 1/\sqrt{1+\gamma\tau_i}$. This yields $O(1/N^2)$ rates for the ergodic gap as well as $\|x^N - \widehat{x}\|^2$.

Several recent works [1, 9–11, 18, 26] have applied inertia and closely-related over-relaxation [13, 15] to the basic method (1.5). For inertia, writing $u^i = (x^i, y^i)$ and $\bar{u}^i = (\bar{x}^i, \bar{y}^i)$, similarly to (1.2), one rebases the algorithm at $\bar{u}^i := (1+\alpha_i)u^i - \alpha_i u^{i-1}$ in place of u^i. In [9], $O(1/N)$ convergence of an *ergodic* gap functional is shown for this method. No $O(1/N^2)$ results are known to us, or results for a non-ergodic gap or iterates. *In this work, we want to improve upon these convergence rate results, possibly by modifying the algorithm.*

A crucial ingredient for inertia to work in (1.2) is a gap unrolling argument. To demonstrate this argument, we take for simplicity $F = 0$. Then (1.2) implies $q^{i+1} := -\tau^{-1}(x^{i+1} - \bar{x}^i) \in \partial G(x^{i+1})$. Defining the auxiliary sequence $\zeta^{i+1} := \lambda_i^{-1}x^{i+1} - (\lambda_i^{-1} - 1)x^i$, for all $\widehat{x} \in X$ one then has[1]

$$C_0 := \frac{1}{2\tau}\|\zeta^0 - \widehat{x}\|^2 \geq -\tau^{-1}\sum_{i=0}^{N-1}\langle \zeta^{i+1} - \zeta^i, \zeta^{i+1} - \widehat{x}\rangle = \sum_{i=0}^{N-1} \lambda_i^{-1}\langle q^{i+1}, \zeta^{i+1} - \widehat{x}\rangle. \tag{1.6}$$

[1]For the inequality apply Pythagoras' identity to convert the right-hand inner product into norms squared. Then cancel repeated terms and estimate the remaining negative terms. The details in abstract form can also be found in Theorem 2.3.

If we do not apply inertia, that is $\lambda_i \equiv 1$, we have $\zeta^{i+1} = x^{i+1}$, so by convexity and Jensen's inequality

$$C_0 \ge \sum_{i=0}^{N-1}\bigl(G(x^{i+1}) - G(\widehat{x})\bigr) \ge N\bigl(G(\tilde{x}^N) - G(\widehat{x})\bigr), \quad \text{where} \quad \tilde{x}^N := \frac{1}{N}\sum_{i=0}^{N-1} x^{i+1}. \tag{1.7}$$

Due to the variable $\tilde{x}^N$, this $O(1/N)$ estimate is ergodic. If, on the other hand, we update λ_i as in (1.3), we can unroll the ergodicity: Since

$$\lambda_i(\zeta^{i+1} - \widehat{x}) = \lambda_i(x^{i+1} - \widehat{x}) + (1-\lambda_i)(x^{i+1} - x^i),$$

we estimate from (1.6) by rearrangements and the definition of the subdifferential that

$$\begin{aligned}
C_0 &\ge \sum_{i=0}^{N-1} \lambda_i^{-2}\Bigl[\lambda_i\langle q^{i+1}, x^{i+1} - \widehat{x}\rangle + (1-\lambda_i)\langle q^{i+1}, x^{i+1} - x^i\rangle\Bigr] \\
&\ge \sum_{i=0}^{N-1} \lambda_i^{-2}\Bigl[\lambda_i(G(x^{i+1}) - G(\widehat{x})) + (1-\lambda_i)(G(x^{i+1}) - G(x^i))\Bigr] \\
&= \sum_{i=0}^{N-1}\Bigl[\lambda_i^{-2}(G(x^{i+1}) - G(\widehat{x})) - (\lambda_i^{-2} - \lambda_i^{-1})(G(x^i) - G(\widehat{x}))\Bigr].
\end{aligned} \tag{1.8a}$$

Telescoping and the recurrence $\lambda_i^{-2} = \lambda_{i+1}^{-2} - \lambda_{i+1}^{-1}$ established from (1.3) now yield

$$C_0 + \lambda_0^{-2}(1-\lambda_0)(G(x^0) - G(\widehat{x})) \ge \lambda_{N-1}^{-2}(G(x^N) - G(\widehat{x})). \tag{1.8b}$$

Since the recurrence also implies that λ_N is of the order $O(1/N^2)$ [3], this gives the improved convergence rate. Similar arguments can be applied to the forward step component F, as we will demonstrate in Section 3 in a more general setting.

How could such argumentation be applied to the PDPS (1.5)? It was discovered in [15] that the method can be written as the "preconditioned proximal point method"

$$0 \in H(u^{i+1}) + W_{i+1}^{-1}M_{i+1}(u^{i+1} - u^i) \tag{1.9}$$

in the space $U := X \times Y$ with the general notation $u = (x, y)$ for the monotone operator $H : U \rightrightarrows U$, the linear preconditioner $M_{i+1} \in \mathcal{L}(U; U)$, and the step length operator $W_{i+1} \in \mathcal{L}(U; U)$ defined as

$$H(u) := \begin{pmatrix} \partial G(x) + K^* y \\ \partial F^*(y) - Kx \end{pmatrix}, \quad M_{i+1} := \begin{pmatrix} I & -\tau_i K^* \\ -\sigma_{i+1}\omega_i K & I \end{pmatrix}, \quad \text{and} \quad W_{i+1} := \begin{pmatrix} \tau_i I & 0 \\ 0 & \sigma_{i+1} I \end{pmatrix}. \tag{1.10}$$

The over-relaxation parameter ω_i and the step length parameters τ_i and σ_{i+1} are as after (1.5). Clearly $H(u) = \partial\hat{G}(u) + \Gamma u$ for the convex function $\hat{G}(u) := G(x) + F^*(y)$ and an anti-symmetric operator Γ. We can thus apply (1.8) to $\hat{G}$. However, the anti-symmetric operator Γ does not arise as a subdifferential, so similar arguments do not apply to it. Indeed, given some inertial parameters $\lambda_i > 0$ and the primal-dual auxiliary sequence $z^{i+1} := \lambda_i^{-1}u^{i+1} - (\lambda_i^{-1} - 1)u^i$, corresponding to ζ^{i+1} above, it does not seem possible to develop a useful estimate out of $\sum_{i=0}^{N-1}\lambda_i^{-1}\langle H(u^{i+1}), z^{i+1} - \widehat{u}\rangle$

alone, unless $\lambda_i \equiv 1$. *In Section 2, we are therefore going to* correct *the inertial scheme against the anti-symmetry of* Γ. We do this in the context of general proximal point methods for the solution of the variational inclusion $0 \in H(\widehat{u})$. We demonstrate how to *test* for convergence rates of such methods based on the ideas introduced in [22] for non-inertial methods.

We adapt and improve the inertial unrolling argument (1.8) to the setting of this testing theory and corrected inertial methods in Section 3. With an eye towards convergence rate proofs, we also develop parameter growth estimates, and briefly demonstrate the theory by application to FISTA. Based on the general results of Sections 2 and 3, we then develop our proposed *inertial, corrected, primal-dual proximal splitting* (IC-PDPS) in Section 4. Using the corrector, we will also be able to incorporate strong convexity based acceleration into the inertial method. We finish with conclusions and numerical experience in Section 5. Readers wishing to simply implement our proposed method, can find it in an explicit and mostly self-contained form near the end in Algorithm 4.1. Only the step length rules need to be taken from a choice of theorems given in the algorithm description.

Notation We write $\overline{\mathbb{R}} := [-\infty, \infty]$ for the extended reals and $\mathcal{L}(X; Y)$ for the space of bounded linear operators between Hilbert spaces X and Y. The identity operator in any space is I. For $T, S \in \mathcal{L}(X; X)$, we write $T \geq S$ when $T - S$ is positive semidefinite. Also for possibly non-self-adjoint T, we introduce the inner product $\langle x, z\rangle_T := \langle Tx, z\rangle$, and, for positive semi-definite T, the semi-norm $\|x\|_T := \sqrt{\langle x, x\rangle_T}$. For a set $A \subset \mathbb{R}$ and a scalar $c \in \mathbb{R}$, we write $A \geq c$ if every element $t \in A$ satisfies $t \geq c$. We write $H : X \rightrightarrows Y$ for H being a set-valued map from X to Y.

2 GENERAL INERTIAL METHODS, CORRECTORS

We will now study the application of inertia to general preconditioned proximal point schemes, of which the PDPS is an instance. As we discussed in the Introduction, [15] showed that the PDPS (1.5) can be written as solving $0 \in H(u^{i+1}) + W_{i+1}^{-1} M_{i+1}(u^{i+1} - u^i)$ for u^{i+1} with the choices (1.10). Further developments in [22, 25] rewrote the method with $\tilde{H}_{i+1} := W_{i+1} H$ as an instance of the more general scheme

$$0 \in \tilde{H}_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - u^i) \tag{PP}$$

that can also model forward steps. This formulation, with the step length operator W_{i+1} uninverted and moved at the front of H turned out to be beneficial for the development of compact convergence rate proofs of the PDPS [22], as well as stochastic extensions that would have W_{i+1} non-invertible [23].

In this section, we will study the application of inertia to (PP). We start by formulating a simple extension of (1.2) to (PP). As in (1.3), writing the inertial parameter as $\alpha_{i+1} = \lambda_{i+1}(\lambda_i^{-1} - 1)$, we now take an invertible linear operator Λ_{i+1} as our fundamental inertial parameter. Given an initial iterate $u^0 = \bar{u}^0 \in U$, we then rebase u^i in (PP) to $\bar{u}^i$ to obtain the method

$$\begin{cases} 0 \in \tilde{H}_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - \bar{u}^i), \\ \bar{u}^{i+1} := u^{i+1} + \Lambda_{i+2}(\Lambda_{i+1}^{-1} - I)(u^{i+1} - u^i). \end{cases} \tag{2.1}$$

We assume that $\tilde{H}_{i+1} : U \rightrightarrows U$ and $M_{i+1}, \Lambda_{i+1} \in \mathcal{L}(U; U)$ on a Hilbert space U.

Remark 2.1. The operator Λ_{i+1} has the index $i+1$ off-by-one compared to λ_i in (1.2) and (1.3). This is for consistency with the historical development of the PDPS (1.5) into the form (1.9) or (PP): compare (1.10), where primal step lengths within the step length operator W_{i+1} have index i, and dual step lengths index $i+1$. This will generally be the case: operator indices agree with dual parameter indices, while primal parameter indices will be one less. We have not reindexed the parameters to maintain the property $\sigma_i\tau_i = \sigma_0\tau_0$ of the PDPS.

We want to correct for any anti-symmetric or otherwise challenging components $\Gamma_{i+1} \in \mathcal{L}(U;U)$ of $\tilde{H}_{i+1}$. We therefore introduce the *corrector*

$$\check{M}_{i+1} := \Gamma_{i+1}(\Lambda_{i+1}^{-1} - I), \tag{2.2}$$

and modify (2.1) into the *general corrected inertial method*

$$\text{(PP-I)} \qquad \begin{cases} 0 \in \tilde{H}_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - \bar{u}^i) + \check{M}_{i+1}(u^{i+1} - u^i), \\ \bar{u}^{i+1} := u^{i+1} + \Lambda_{i+2}(\Lambda_{i+1}^{-1} - I)(u^{i+1} - u^i). \end{cases}$$

In this section, our task is to develop general convergence estimates for (PP-I), which we will then use to prove convergence rates of more specific instances of the general method, in particular the inertial, corrected, PDPS in Section 4. To interpret the main assumption of our abstract convergence estimate, and for later use, we recall the following three-point inequality:

Lemma 2.2. Let $F : X \to \overline{\mathbb{R}}$ be proper, convex, lower semicontinuous with ∇F L-Lipschitz. Then

$$\langle \nabla F(z), x - \widehat{x} \rangle \geq F(x) - F(\widehat{x}) - \frac{L}{2}\|x - z\|^2 \quad (\widehat{x}, z, x \in X).$$

Proof. Since F has L-Lipschitz gradient, it is smooth in the sense of convex analysis (also known as satisfying the *descent inequality*), $F(z) - F(x) \geq \langle \nabla F(z), z - x \rangle - \frac{L}{2}\|x - z\|^2$. By convexity $F(\widehat{x}) - F(z) \geq \langle \nabla F(z), \widehat{x} - z \rangle$. Summing these two estimates, we obtain the claim. □

We will develop our convergence estimates following the testing framework of [22]. The idea introduced there was to pick a suitably designed testing operator $Z_{i+1} \in \mathcal{L}(U;U)$, and then apply the testing functional $u \mapsto \langle u, u^{i+1} - \widehat{u} \rangle_{Z_{i+1}}$ to both sides of (PP). An almost trivial argument based on a simple assumption on $\tilde{H}_{i+1}$ and Pythagoras' (three-point) identity would then show that $Z_{i+1}M_{i+1}$ forms a local metric that measures convergence rates. However, presently, we cannot in general obtain estimates on the principal sequence $\{u^i\}_{i\in\mathbb{N}}$. Rather, we will obtain estimates on the auxiliary sequence $\{z^i\}_{i\in\mathbb{N}}$, defined through

$$z^0 := u^0 \quad \text{and} \quad z^{i+1} := \Lambda_{i+1}^{-1}u^{i+1} - (\Lambda_{i+1}^{-1} - I)u^i \quad (i \in \mathbb{N}). \tag{2.3}$$

This adds some additional complexity to the main condition (2.4) of the next theorem. We will motivate the condition after the proof.

Theorem 2.3. On a Hilbert space U, for $i = 0, \ldots, N-1$, let $\tilde{H}_{i+1} : U \rightrightarrows U$ as well as $M_{i+1}, Z_{i+1}, \Gamma_{i+1}, \Lambda_i, \in \mathcal{L}(U;U)$ with Λ_i invertible. Given an initial iterate $u^0 = \bar{u}^0 \in U$, let $\{u^{i+1}\}_{i\in\mathbb{N}}$ be defined through the solution of (PP-I), and the auxiliary sequence $\{z^i\}_{i\in\mathbb{N}}$ by (2.3). Suppose, for $i = 1, \ldots, N-1$, *that $Z_{i+1}M_{i+1} \ge 0$ is self-adjoint, and for some $\widehat{u} \in U$ and a placeholder real value $\mathcal{V}_{i+1}(\widehat{u}) \in \mathbb{R}$ that*

$$\langle \tilde{H}_{i+1}(u^{i+1}) - \Gamma_{i+1}(u^{i+1} - \widehat{u}), z^{i+1} - \widehat{u}\rangle_{\Lambda_{i+1}^* Z_{i+1}} \ \ge\ \mathcal{V}_{i+1}(\widehat{u}) - \frac{1}{2}\|z^{i+1} - z^i\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1} \Lambda_{i+1}} \tag{2.4}$$

and

$$\Lambda_{i+1}^* Z_{i+1}(M_{i+1}\Lambda_{i+1} + 2\Gamma_{i+1}) \ge \Lambda_{i+2}^* Z_{i+2} M_{i+2}\Lambda_{i+2}. \tag{2.5}$$

Then

$$\frac{1}{2}\|z^N - \widehat{u}\|^2_{\Lambda_{N+1}^* Z_{N+1} M_{N+1}\Lambda_{N+1}} + \sum_{i=0}^{N-1} \mathcal{V}_{i+1}(\widehat{u}) \le \frac{1}{2}\|z^0 - \widehat{u}\|^2_{\Lambda_1^* Z_1 M_1 \Lambda_1} \quad (N \ge 1). \tag{2.6}$$

Proof. Application of $\langle\,\cdot\,, Z_{i+1}^*\Lambda_{i+1}(z^{i+1} - \widehat{u})\rangle$ to the main inclusion of (PP-I) yields for some $q^{i+1} \in \tilde{H}_{i+1}(u^{i+1})$ that

$$0 = \langle q^{i+1} + M_{i+1}(u^{i+1} - \bar{u}^i) + \check{M}_{i+1}(u^{i+1} - u^i), \Lambda_{i+1}(z^{i+1} - \widehat{u})\rangle_{Z_{i+1}}. \tag{2.7}$$

By (2.3) and (2.2), we deduce that

$$\check{M}_{i+1}(u^{i+1} - u^i) = \Gamma_{i+1}(\Lambda_{i+1}^{-1} - I)(u^{i+1} - u^i) = \Gamma_{i+1}(z^{i+1} - u^{i+1}).$$

By the definition $\bar{u}^i$ in (PP-I), and of the auxiliary sequence $\{z^{i+1}\}_{i\in\mathbb{N}}$ in (2.3), taking $u^{-1} := u^0$, moreover,

$$\begin{aligned}\Lambda_{i+1}(z^{i+1} - z^i) &= u^{i+1} - (I - \Lambda_{i+1})u^i - \Lambda_{i+1}[\Lambda_i^{-1}u^i - (\Lambda_i^{-1} - I)u^{i-1}] \\ &= u^{i+1} - [I - \Lambda_{i+1} + \Lambda_{i+1}\Lambda_i^{-1}]u^i - \Lambda_{i+1}(I - \Lambda_i^{-1})u^{i-1} \\ &= u^{i+1} - \bar{u}^i.\end{aligned} \tag{2.8}$$

Therefore, we transform (2.7) into

$$0 = \langle q^{i+1} + M_{i+1}\Lambda_{i+1}(z^{i+1} - z^i) + \Gamma_{i+1}(z^{i+1} - u^{i+1}), \Lambda_{i+1}(z^{i+1} - \widehat{u})\rangle_{Z_{i+1}}. \tag{2.9}$$

Writing for brevity $A := \Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}$, which by assumption is self-adjoint and positive semi-definite, the standard three-point formula or Pythagoras' identity states

$$\langle z^{i+1} - z^i, z^{i+1} - \widehat{u}\rangle_A = \frac{1}{2}\|z^{i+1} - z^i\|_A^2 - \frac{1}{2}\|z^i - \widehat{u}\|_A^2 + \frac{1}{2}\|z^{i+1} - \widehat{u}\|_A^2.$$

We therefore transform (2.9) into

$$\begin{aligned}0 = \langle q^{i+1} + \Gamma_{i+1}(z^{i+1} - u^{i+1}), z^{i+1} - \widehat{u}\rangle_{\Lambda_{i+1}^* Z_{i+1}} &+ \frac{1}{2}\|z^{i+1} - z^i\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}} \\ - \frac{1}{2}\|z^i - \widehat{u}\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}} &+ \frac{1}{2}\|z^{i+1} - \widehat{u}\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}}.\end{aligned}$$

Using (2.4), we obtain

$$\begin{aligned}0 \ge \mathcal{V}_{i+1}(\widehat{u}) &- \frac{1}{2}\|z^i - \widehat{u}\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}} \\ &+ \frac{1}{2}\|z^{i+1} - \widehat{u}\|^2_{\Lambda_{i+1}^* Z_{i+1} M_{i+1}\Lambda_{i+1}} + \langle \Gamma_{i+1}(z^{i+1} - \widehat{u}), z^{i+1} - \widehat{u}\rangle_{\Lambda_{i+1}^* Z_{i+1}}.\end{aligned}$$

Using (2.5) and summing over $i = 0, \dots, N-1$ establishes (2.6). ◻

Remark 2.4. *Consider $\Lambda_{i+1} = Z_{i+1} = M_{i+1} = I$. Then (2.4) reads*

$$\langle \tilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u} \rangle \geq \mathcal{V}_{i+1}(\widehat{u}) + \langle u^{i+1} - \widehat{u}, u^{i+1} - \widehat{u} \rangle_{\Gamma_{i+1}} - \frac{1}{2}\|u^{i+1} - u^i\|^2. \tag{2.10}$$

With $\tau L \leq 1$ and $\Gamma_{i+1} = 0$, take first $\tilde{H}_{i+1}(u) = \tau \nabla F(u^i)$, fixing ∇F to be evaluated at the previous iteration to obtain a gradient descent method. Then it is easy to see how this estimate with $\mathcal{V}_{i+1}(\widehat{u}) = \tau[F(u^{i+1}) - F(\widehat{u})]$ follows from Lemma 2.2. Thus $\mathcal{V}_{i+1}$ measures function value differences. Similarly, if $\tilde{H}_{i+1}(u) = \tau \partial G(u)$ for non-smooth but (γ-strongly) convex G, we can take $\Gamma_{i+1} = \tau\gamma I$ in (2.10). We can also combine $\tilde{H}_{i+1}(u) = \tau[\partial G(u) + \nabla F(u^i)]$ to obtain forward–backward splitting. In other words, (2.4) is an operator-relative inertia-aware convexity and smoothness condition, where the variable $\mathcal{V}_{i+1}(\widehat{u})$ can be used to model function value and other gap estimates. We will need this operator-relativity to apply distinct inertial and testing parameters on the primal and dual variables of the PDPS; compare the block structure of (1.10).

Minding this interpretation of (2.4), the claim (2.6) of the theorem with additional positivity and growth assumptions can thus be used to show the convergence of the sum of the value estimates $\mathcal{V}_{i+1}(\widehat{u})$ to zero, as well as $z^N \to \widehat{u}$. Our task in the following is to obtain that growth, and to unroll the sum into a simple estimate.

3 UNROLLING AND PARAMETER GROWTH ESTIMATES

To develop the inertial, corrected, primal-dual proximal splitting method, we will seek to satisfy the conditions of Theorem 2.3 for an algorithm inspired by the proximal point interpretation of the PDPS. Before we do this, in this section, we will prove general inertial unrolling arguments (Section 3.2), refining (1.8) to the testing framework. We also prove parameter growth estimates with an eye towards converge rate proofs (Section 3.3), and demonstrate how our corrector term allows combining inertia with strong convexity-based acceleration (3.4). We finish by applying these estimates to the FISTA to demonstrate how it fits into our overall approach (Section 3.4). This also demonstrates how our approach works without the additional challenges of the primal-dual setup.

3.1 SCALAR PARAMETER CHOICES

To place the general estimates that make up the major part of this section into context, we start by specialising Theorem 2.3 to $\Lambda_{i+1} = \lambda_i I$, $Z_{i+1} = \phi_i I$, $W_{i+1} = \tau_i I$, $M_{i+1} = I$, and $\Gamma_{i+1} := \gamma\tau_i I$ for some scalars $\lambda_i, \phi_i, \tau_i > 0$ and $\gamma \geq 0$. Also taking $\tilde{H}_{i+1}(x) := \tau_i(\partial G(x) + \nabla F(x^i))$, we immediately rewrite (PP-I), with change of symbol[2] u into x, as

$$\begin{cases} 0 \in \tau_i[\partial G(x^{i+1}) + \nabla F(x^i)] + (x^{i+1} - \bar{x}^i) + \gamma\tau_i(\lambda_i^{-1} - 1)(x^{i+1} - x^i), \\ \bar{x}^{i+1} := x^{i+1} + \lambda_{i+1}(\lambda_i^{-1} - 1)(x^{i+1} - x^i). \end{cases} \tag{PP-i}$$

Moreover, with the change of symbol[2] of z into ζ, the auxiliary sequence defined in (2.3) becomes

$$\zeta^0 := x^0 \quad \text{and} \quad \zeta^{i+1} := \lambda_i^{-1} x^{i+1} - (\lambda_i^{-1} - 1)x^i \quad (i \in \mathbb{N}). \tag{3.1}$$

Immediately, Theorem 2.3 specialises into:

Corollary 3.1. *On a Hilbert space X, let $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous, with F differentiable. Let $\lambda_i, \phi_i, \tau_i > 0$, $(i = 0, \ldots, N-1)$, and $\gamma \geq 0$. For an initial iterate $x^0 = \bar{x}^0 \in X$, let $\{x^{i+1}\}_{i=0}^{N-1}$ be generated by (PP-i), and the auxiliary sequence $\{\zeta^i\}_{i=0}^{N-1}$ by (3.1). For each $i = 0, \ldots, N-1$, suppose for some $\widehat{x} \in X$ and a placeholder value $\mathcal{V}_{i+1}(\widehat{x}) \in \mathbb{R}$, we have the estimate*

$$\phi_i\lambda_i\tau_i\langle \partial G(x^{i+1}) + \nabla F(x^i) - \gamma(x^{i+1} - \widehat{x}), \zeta^{i+1} - \widehat{x}\rangle \geq \mathcal{V}_{i+1}(\widehat{x}) - \frac{\lambda_i^2\phi_i}{2}\|\zeta^{i+1} - \zeta^i\|^2 \tag{3.2}$$

and the inequality

$$\lambda_{i+1}^2\phi_{i+1} \leq \lambda_i^2\phi_i(1 + 2\gamma\lambda_i^{-1}\tau_i). \tag{3.3}$$

Then

$$\frac{\phi_N\lambda_N^2}{2}\|\zeta^N - \widehat{x}\|^2 + \sum_{i=0}^{N-1}\mathcal{V}_{i+1}(\widehat{x}) \leq \frac{\phi_0\lambda_0^2}{2}\|\zeta^0 - \widehat{x}\|^2 \quad (N \geq 1). \tag{3.4}$$

3.2 Inertial unrolling

We start by refining the proximal step inertial unrolling argument (1.8). We will in Section 3.3 see that the recurrence inequality (3.5) assumed by the next lemma generalises the recurrence $\lambda_i^{-2} = \lambda_{i+1}^{-2} - \lambda_{i+1}^{-1}$ from the Introduction, satisfied by the FISTA.

Lemma 3.2. *Let $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose $\lambda_i \in (0, 1]$ and $\phi_i, \tau_i > 0$ satisfy the recurrence inequality*

$$\phi_{i+1}\tau_{i+1}(1 - \lambda_{i+1}) \leq \phi_i\tau_i \quad (i = 0, \ldots, N-1). \tag{3.5}$$

For any given $\{x^i\}_{i=0}^N$, let the auxiliary variables $\{\zeta^i\}_{i=0}^N$ be generated by (3.1). Assume $\partial G(x^{i+1})$ to be non-empty for $i = 0, \ldots, N-1$, and $\widehat{x} \in [\partial G]^{-1}(0)$. Then

$$\begin{aligned}\sum_{i=0}^{N-1}\inf_{q^{i+1}\in\partial G(x^{i+1})}\phi_i\tau_i\lambda_i\langle q^{i+1}, \zeta^{i+1} - \widehat{x}\rangle \geq{}& \phi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) \\ &- \phi_0\tau_0(1 - \lambda_0)(G(x^0) - G(\widehat{x})).\end{aligned} \tag{3.6}$$

Proof. For all $i = 0, \ldots, N-1$, pick $q^{i+1} \in \partial G(x^{i+1})$, and define

$$s_N^G = \sum_{i=0}^{N-1}\phi_i\tau_i\lambda_i\langle q^{i+1}, \zeta^{i+1} - \widehat{x}\rangle.$$

Then we need to show that

$$s_N^G \geq \phi_{N-1}\tau_{N-1}(G(x^N) - G(\widehat{x})) - \phi_0\tau_0(1 - \lambda_0)(G(x^0) - G(\widehat{x})). \tag{3.7}$$

[2] We reserve the symbols u and z for the abstract (Section 2) and primal-dual (Section 4) problems. In the latter we take primal-dual pairs $u = (x, y)$ and $z = (\zeta, \eta)$, so the primal variables match the symbols of this section.

Observe that the auxiliary variables $\{\zeta^{i+1}\}_{i=0}^{N-1}$ satisfy

$$\lambda_i(\zeta^{i+1} - \widehat{x}) = \lambda_i(x^{i+1} - \widehat{x}) + (1-\lambda_i)(x^{i+1} - x^i). \tag{3.8}$$

With this and the convexity of G, we estimate

$$\begin{aligned}
s_N^G &= \sum_{i=0}^{N-1} \phi_i\tau_i\Big[\lambda_i\langle q^{i+1}, x^{i+1} - \widehat{x}\rangle + (1-\lambda_i)\langle q^{i+1}, x^{i+1} - x^i\rangle\Big] \\
&\geq \sum_{i=0}^{N-1} \phi_i\tau_i\Big[\lambda_i(G(x^{i+1}) - G(\widehat{x})) + (1-\lambda_i)(G(x^{i+1}) - G(x^i))\Big] \\
&= \sum_{i=0}^{N-1}\Big[\phi_i\tau_i(G(x^{i+1}) - G(\widehat{x})) - \phi_i\tau_i(1-\lambda_i)(G(x^i) - G(\widehat{x}))\Big].
\end{aligned} \tag{3.9}$$

Since $G(x^i) \geq G(\widehat{x})$, the recurrence inequality (3.5) together with a telescoping argument now give (3.7). ◻

We can also include a forward step in the unrolling argument:

Lemma 3.3. *Let $G, F : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. Suppose F has L-Lipschitz gradient, and that $\lambda_i \in (0,1]$ and $\phi_i, \tau_i > 0$ satisfy the recurrence inequality (3.5) for $i = 0, \ldots, N-1$. For any given $\{x^i\}_{i=0}^N$, let the auxiliary variables $\{\zeta^i\}_{i=0}^N$ be generated by (3.1). Assume $\partial G(x^{i+1})$ to be non-empty for all $i = 0, \ldots, N-1$, and that $\widehat{x} \in [\partial G + \nabla F]^{-1}(0)$. Then*

$$\begin{aligned}
&\sum_{i=0}^{N-1} \inf_{q^{i+1}\in\partial G(x^{i+1})} \left[\phi_i\tau_i\lambda_i\langle q^{i+1} + \nabla F(\bar{x}^i), \zeta^{i+1} - \widehat{x}\rangle + \frac{\phi_i\tau_i\lambda_i^2 L}{2}\|\zeta^{i+1} - \zeta^i\|^2\right] \\
&\qquad \geq \phi_{N-1}\tau_{N-1}[(G+F)(x^N) - (G+F)(\widehat{x})] - \phi_0\tau_0(1-\lambda_0)[(G+F)(x^0) - (G+F)(\widehat{x})].
\end{aligned} \tag{3.10}$$

Proof. Similarly to (2.8), $\lambda_i(\zeta^{i+1} - \zeta^i) = (x^{i+1} - \bar{x}^i)$. By (3.8) and Lemma 2.2, therefore

$$\begin{aligned}
s_N^F &:= \sum_{i=0}^{N-1}\left[\phi_i\tau_i\lambda_i\langle\nabla F(\bar{x}^i), \zeta^{i+1} - \widehat{x}\rangle + \frac{\phi_i\tau_i\lambda_i^2 L}{2}\|\zeta^{i+1} - \zeta^i\|^2\right] \\
&= \sum_{i=0}^{N-1}\phi_i\tau_i\left[\lambda_i\langle\nabla F(\bar{x}^i), x^{i+1} - \widehat{x}\rangle + (1-\lambda_i)\langle\nabla F(\bar{x}^i), x^{i+1} - x^i\rangle + \frac{L}{2}\|x^{i+1} - \bar{x}^i\|^2\right] \\
&\geq \sum_{i=0}^{N-1}\phi_i\tau_i\left[\lambda_i(F(x^{i+1}) - F(\widehat{x})) + (1-\lambda_i)(F(x^{i+1}) - F(x^i))\right] \\
&= \sum_{i=0}^{N-1}\left[\phi_i\tau_i(F(x^{i+1}) - F(\widehat{x})) - \phi_i\tau_i(1-\lambda_i)(F(x^i) - F(\widehat{x}))\right].
\end{aligned}$$

Picking $q^{i+1} \in \partial G(x^{i+1})$, $(i = 0, \ldots, N-1)$, and summing with the estimate (3.9) for G, we deduce

$$s_N^G + s_N^F \geq \sum_{i=0}^{N-1}\phi_i\tau_i\left[[(G+F)(x^{i+1}) - (G+F)(\widehat{x})] - (1-\lambda_i)[(G+F)(x^i) - (G+F)(\widehat{x})]\right].$$

Since $(G+F)(x^i) \geq (G+F)(\widehat{x})$, the recurrence inequality (3.5) together with a telescoping argument now give the claim. ◻

3.3 PARAMETER GROWTH ESTIMATES

As suggested by the unrolled estimates (3.6) and (3.10), we want to make $\phi_{N-1}\tau_{N-1}$ grow as fast as possible while satisfying (3.5) and $\lambda_i \in (0, 1]$. We now develop such estimates through a series of lemmas. The first of these lemmas with $\epsilon = 0$ is the FISTA rate argument [3, Lemma 4.3].

Lemma 3.4. *With $\lambda_0 = 1$, suppose $\lambda_i^{-2} - \epsilon\lambda_i^{-1} = \lambda_{i+1}^{-2} - \lambda_{i+1}^{-1}$ for some $\epsilon \in [-1, 1]$ and all $i = 0, \ldots, N-1$. Then $\{\lambda_i^{-1}\}_{i\in\mathbb{N}}$ is non-decreasing, $\lambda_N^{-1} \geq 1 + (1-\epsilon)N/2$, and we equivalently define λ_{i+1} through*

$$\lambda_{i+1} = \frac{2}{1 + \sqrt{1 + 4(\lambda_i^{-2} - \epsilon\lambda_i^{-1})}}. \tag{3.11}$$

Proof. The update (3.11) is a simple solution of the quadratic equation $\lambda_i^{-2} - \epsilon\lambda_i^{-1} = \lambda_{i+1}^{-2} - \lambda_{i+1}^{-1}$. The latter also rearranges as

$$\lambda_{i+1}^{-2} - \lambda_{i+1}^{-1} = \lambda_i^{-2} - \lambda_i^{-1} + (1-\epsilon)\lambda_i^{-1}. \tag{3.12}$$

This shows that $\{t_i := \lambda_i^{-2} - \lambda_i^{-1}\}_{i\in\mathbb{N}}$ is non-decreasing. Since $\lambda_i^{-1} = \frac{1}{2}(1 + \sqrt{1 + 4t_i})$, we obtain the claim that $\{\lambda_i^{-1}\}_{i\in\mathbb{N}}$ is non-decreasing.

We still need to prove the rate-of-growth claim. Since $\lambda_0 = 1$, (3.12) also yields

$$\lambda_N^{-2} - \lambda_N^{-1} = \sum_{i=0}^{N-1}(1-\epsilon)\lambda_i^{-1}.$$

Let us make the inductive assumption that $\lambda_i^{-1} \geq 1 + (1-\epsilon)i/2$ for $i = 0, \ldots, N-1$. Clearly this holds for $N = 0$ by the choice $\lambda_0 = 1$, taking care of the inductive base. For the inductive step, we get from above and $\epsilon \in [-1, 1]$ that

$$\lambda_N^{-2} - \lambda_N^{-1} \geq (1-\epsilon)N + \frac{(1-\epsilon)^2}{4}N(N-1) \geq \frac{1-\epsilon}{2}N + \left(\frac{1-\epsilon}{2}\right)^2 N^2.$$

This quadratic inequality together with $\lambda_N > 0$ imply

$$\lambda_N^{-1} \geq \frac{1 + \sqrt{1 + 2(1-\epsilon)N + (1-\epsilon)^2N^2}}{2} = 1 + \frac{1-\epsilon}{2}N,$$

which verifies the inductive step and establishes the claim. □

Lemma 3.5. *The sequence $\{\phi_i\tau_i\}_{i\in\mathbb{N}}$ is non-decreasing and the conditions (3.3) and (3.5) hold, more precisely*

$$\lambda_{i+1}^2\phi_{i+1} = \lambda_i^2\phi_i(1 + 2\gamma\lambda_i^{-1}\tau_i) \quad \text{and} \quad \phi_{i+1}\tau_{i+1}(1-\lambda_{i+1}) = (1-\epsilon\lambda_i)\phi_i\tau_i \tag{3.13}$$

for all $i \in \mathbb{N}$ for some $\epsilon \in [0, 1)$ in the following cases:

(i) If $\gamma = 0$ and we take $\tau_i \equiv \tau$ for any $\tau > 0$; $\phi_i := \lambda_i^{-2}$; $\phi_0 = \lambda_0 = 1$, and update λ_{i+1} for any $\epsilon \in [0, 1]$ according to (3.11). Then also

$$\phi_N\tau_N \geq (1-\epsilon)^2N^2\tau/4 \quad \text{and} \quad \lambda_N^2\phi_N = 1 \quad (N \in \mathbb{N}).$$

(ii) *If $\gamma > 0$ and we take $\lambda_i \equiv \lambda \in (0,1)$ and $\tau_i \equiv \tau := \lambda^2/[2\gamma(1-\lambda)]$ constants, and $\phi_{i+1} = c\phi_i$ with $c := (1-\epsilon\lambda)/(1-\lambda) > 1$ for any $\epsilon \in [0,1)$ and $\phi_0 > 0$. Then also*

$$\phi_N\tau_N \geq \phi_0\tau c^N \quad \textit{and} \quad \lambda_N^2\phi_N \geq \lambda^2\phi_0 c^N \quad (N \in \mathbb{N}).$$

(iii) *If we are constrained to have $\phi_i = c_0\tau_i^{-2}$ for some constant $c_0 > 0$, and with $\lambda_0 = 1$, $\tau_0 > 0$ and $\epsilon \in [0,1)$ update*

$$\tau_{i+1} = \frac{1-\lambda_{i+1}}{1-\epsilon\lambda_i}\tau_i \quad \textit{and} \quad \lambda_{i+1} = \frac{\sqrt{\lambda_i^2 + 2\gamma\lambda_i\tau_i}}{1-\epsilon\lambda_i + \sqrt{\lambda_i^2 + 2\gamma\lambda_i\tau_i}}. \tag{3.14}$$

Then, for some constants $c, c' > 0$, for all $N \in \mathbb{N}$, also

$$\begin{aligned} &\phi_N\tau_N \geq c'N^2 && \textit{and} && \lambda_N^2\phi_N \geq cN^2 && (\gamma > 0), \\ &\phi_N\tau_N \geq (1-\epsilon)\tau_0^{-1}N && \textit{and} && \lambda_N^2\phi_N = c_0\tau_0^{-2} && (\gamma = 0). \end{aligned}$$

The choice $\epsilon = 0$ in (iii) would be the simplest, and also optimal in the sense that both (3.3) and (3.5) would hold as equalities. However, we will see that a non-zero choice performs significantly better in practise—with the same asymptotic guarantees.

Proof. It is clear that the inequalities (3.3) and (3.5) follow from (3.13) and $\epsilon \geq 0$.

(i) Since $\gamma = 0$, the first part of (3.13) holds when $\phi_i\lambda_i^2 = \phi_0\lambda_0^2$. This follows from our choices $\phi_0 = \lambda_0 = 1$ and $\phi_i = \lambda_i^{-2}$. Inserting $\phi_i = \lambda_i^{-2}$ and $\tau_i \equiv \tau$, the second part of (3.13) reduces to $\lambda_i^{-2} - \epsilon\lambda_i^{-1} = \lambda_{i+1}^{-2} - \lambda_{i+1}^{-1}$. Lemma 3.4 shows that $\phi_N\tau_N = \lambda_N^{-2}\tau \geq \tau(1-\epsilon)^2N^2/4$. This is the claimed estimate. Moreover, since $\{\lambda_i^{-1}\}_{i\in\mathbb{N}}$ is non-decreasing by Lemma 3.4, we see that $\{\phi_i\tau_i\}_{i\in\mathbb{N}}$ is non-decreasing.

(ii) The second part of (3.13) agrees with the chosen update rule for ϕ_{i+1}. Inserting this rule into the first part of (3.13) and using the fact that also $\tau_i \equiv \tau$, we see the latter to be satisfied if $(1-\lambda)(1+2\gamma\lambda^{-1}\tau) = 1$. This is satisfied by our chosen $\tau = \lambda^2/[2\gamma(1-\lambda)]$. Since τ and λ are constants, the claimed growth estimates follow from $\phi_N = \phi_0 c^N$. Clearly $\{\phi_i\tau_i = \phi_0\tau c^i\}_{i\in\mathbb{N}}$ is non-decreasing.

(iii) Finally, with $\phi_i = c_0\tau_i^{-2}$, (3.13) holds if

$$\lambda_{i+1}^2\tau_{i+1}^{-2} = \lambda_i^2\tau_i^{-2}(1+2\gamma\lambda_i^{-1}\tau_i) \quad \text{and} \quad \tau_{i+1}^{-1}(1-\lambda_{i+1}) = (1-\epsilon\lambda_i)\tau_i^{-1}. \tag{3.15}$$

The latter agrees with our update rule for τ_{i+1}. Inserting this, the former holds if $\lambda_{i+1}(1-\epsilon\lambda_i) = (1-\lambda_{i+1})\sqrt{\lambda_i^2 + 2\gamma\lambda_i\tau_i}$. This is satisfied by our update rule for λ_{i+1}.

To derive the growth estimates, suppose first that $\gamma > 0$. With $\theta_i := \lambda_i\tau_i^{-1} = c_0^{-1/2}\lambda_i\phi_i^{1/2}$, the first part of (3.15) reads $\theta_{i+1}^2 = \theta_i^2(1+2\gamma\theta_i)$. Thus $\{\theta_i\}_{i\in\mathbb{N}}$ is non-decreasing, moreover, the recurrence is of the same form as the standard acceleration rule for the PDPS, where we would have $\phi_i = c_0\tau_i^{-2}$ in place of θ_i; compare Section 1 and [7, 25]. Hence $\theta_i \geq ci$ for some constant $c > 0$. Since $\lambda_i^2\phi_i = c_0\theta_i^2$, this gives one of the claimed rates. From the second part of (3.15),

$$\tau_{i+1}^{-1} = (1-\epsilon\lambda_i)\tau_i^{-1} + \lambda_{i+1}\tau_{i+1}^{-1} = \tau_i^{-1} - \epsilon\theta_i + \theta_{i+1}. \tag{3.16}$$

Repeating this recursively, since $\theta_0 = \tau_0^{-1}$, for some constant $c' > 0$,

$$\tau_N^{-1} = \tau_0^{-1} - \epsilon\theta_0 + \theta_N + \sum_{i=1}^{N-1}(1-\epsilon)\theta_i \geq (1-\epsilon)\tau_0^{-1} + cN + \sum_{i=0}^{N-1}(1-\epsilon)ci \geq c'N^2. \tag{3.17}$$

Therefore also $\phi_N\tau_N = c_0\tau_N^{-1}$ has the claimed growth estimate.

Suppose then that $\gamma = 0$. We obtain (3.16) as above, however now with constant $\theta_i \equiv \lambda_i\tau_i^{-1} = \lambda_0\tau_0^{-1} = \tau_0^{-1}$. Therefore, arguing as in (3.17) only yields $\tau_N^{-1} \geq (1-\epsilon)\tau_0^{-1}N$. This is again the claimed growth estimate.

Finally, since $\{\theta_i\}_{i\in\mathbb{N}}$ is in both cases ($\gamma = 0$ or $\gamma > 0$) non-decreasing, we see from (3.16) that $\{\tau_i^{-1}\}_{i\in\mathbb{N}}$ and consequently $\{\phi_i\tau_i = c_0\tau_i^{-1}\}_{i\in\mathbb{N}}$ are non-decreasing. □

3.4 COMBINING INERTIA WITH STRONG CONVEXITY

Let $G : X \to \overline{\mathbb{R}}$ be proper, lower semicontinuous, and (strongly) convex with parameter $\gamma \geq 0$. We now demonstrate with a simple inertial proximal point method, (PP-i) with $F = 0$, how we are able to incorporate strong convexity based acceleration with inertia. We do this by considering the convex functions

$$G_\gamma(x;\widehat{x}) := G(x) - \frac{\gamma}{2}\|x - \widehat{x}\|^2. \tag{3.18}$$

Indeed, $0 \in \partial G(\widehat{x})$ if and only if $0 \in \partial G_\gamma(\widehat{x};\widehat{x})$ with $\partial G_\gamma(x;\widehat{x}) = \partial G(x) - \gamma(x - \widehat{x})$. Lemma 3.2 applied to $G_\gamma(\,\cdot\,;\widehat{x})$ thus shows

$$\sum_{i=0}^{N-1}\mathcal{V}_{i+1}(\widehat{x}) \geq \phi_{N-1}\tau_{N-1}[G_\gamma(x^N;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] - \phi_0\tau_0(1-\lambda_0)[G_\gamma(x^0;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] \tag{3.19}$$

for

$$\mathcal{V}_{i+1}(\widehat{x}) := \inf_{q^{i+1}\in\partial G(x^{i+1})} \phi_i\tau_i\lambda_i\langle q^{i+1} - \gamma(x^{i+1} - \widehat{x}), \zeta^{i+1} - \widehat{x}\rangle.$$

This choice of $\mathcal{V}_{i+1}(\widehat{x})$ by definition verifies (3.2). It therefore follows from Corollary 3.1 with $F \equiv 0$ that

$$\begin{aligned}\frac{\phi_N\lambda_N^2}{2}\|\zeta^N - \widehat{x}\|^2 &+ \phi_{N-1}\tau_{N-1}[G_\gamma(x^N;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] \\ &\leq \frac{\phi_0\lambda_0^2}{2}\|\zeta^0 - \widehat{x}\|^2 + \phi_0\tau_0(1-\lambda_0)[G_\gamma(x^0;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] \quad (N \geq 1).\end{aligned}$$

If we choose our parameters according to Lemma 3.5 (ii) , then $\phi_N\tau_N$ and $\lambda_N^2\phi_N$ grow exponentially. Crucially $G_\gamma(x^N;\widehat{x}) \geq G_\gamma(\widehat{x};\widehat{x})$, so this implies linear convergence of $G_\gamma(x^N;\widehat{x}) \to G_\gamma(\widehat{x};\widehat{x})$ and of the auxiliary sequence $\zeta^N \to \widehat{x}$.

3.5 CONNECTIONS

We now discuss how our results relate to known algorithms.

Example 3.1 (FISTA). Let $G : X \to \overline{\mathbb{R}}$ and $F : X \to \mathbb{R}$ be convex, proper, and lower semi-continuous with ∇F existing and L-Lipschitz. If $\gamma = 0$, take $\tau_i \equiv \tau \in (0, 1/L]$, and λ_{i+1} by (3.11) for $\lambda_0 = 1$ and $\epsilon = 0$. Then given initial iterates $\bar{x}^0 = x^0$, (PP-i) becomes the inertial forward–backward splitting or FISTA

$$\begin{cases} x^{i+1} := \operatorname{prox}_{\tau G}(\bar{x}^i - \tau \nabla F(\bar{x}^i)), \\ \bar{x}^{i+1} := x^{i+1} + \lambda_{i+1}(\lambda_i^{-1} - 1)(x^{i+1} - x^i). \end{cases}$$

We have $G(x^N) + F(x^N) \to G(\widehat{x}) + F(\widehat{x})$ at the rate $O(1/N^2)$ for any minimiser $\widehat{x}$ of $G + F$.

Demonstration. The algorithm is clear from (PP-i). Lemma 3.5(i) shows that (3.3) and (3.5) are satisfied, $\phi_N \tau_N \geq \tau N^2/4$, and $\lambda_N^2 \phi_N = 1$. Clearly (3.2) holds with

$$\mathcal{V}_{i+1}(\widehat{x}) := \inf_{q^{i+1} \in \partial G(x^{i+1})} \phi_i \tau_i \lambda_i \langle q^{i+1} + \nabla F(\bar{x}^i) - \gamma(x^{i+1} - \widehat{x}), \zeta^{i+1} - \widehat{x} \rangle + \frac{\lambda_i^2 \phi_i}{2} \|\zeta^{i+1} - \zeta^i\|^2.$$

Using in (3.10) the fact that $\tau_i L \leq 1$, similarly to Section 3.4, we may therefore refer to Lemma 3.3 and Corollary 3.1 to verify the estimate

$$\begin{aligned} \text{(3.20)} \quad & \frac{\phi_N \lambda_N^2}{2} \|\zeta^N - \widehat{x}\|^2 + \phi_{N-1} \tau_{N-1} [(\hat{G}_\gamma + F)(x^N) - (\hat{G}_\gamma + F)(\widehat{x})] \\ & \qquad \leq \frac{\phi_0 \lambda_0^2}{2} \|\zeta^0 - \widehat{x}\|^2 + \phi_0 \tau_0 (1 - \lambda_0) [(\hat{G}_\gamma + F)(x^0) - (\hat{G}_\gamma + F)(\widehat{x})], \end{aligned}$$

where we use the short-hand notation $\hat{G}_\gamma := G_\gamma(\,\cdot\,; \widehat{x})$. Inserting our choice $\lambda_0 = 1$ with $\gamma = 0$, $\zeta^0 = x^0$ and the growth estimates from above, we obtain

$$(G + F)(x^N) \leq (G + F)(\widehat{x}) + \frac{2}{(N-1)^2 \tau} \|x^0 - \widehat{x}\|^2 \quad (N \geq 2).$$

This verifies the claimed convergence rates. □

Example 3.2 (FISTA combined with strong convexity). In Example 3.1, suppose in addition that G is strongly convex with parameter $\gamma > 0$. Take $0 < \lambda \leq \sqrt{L^{-2}\gamma^2 + 2L^{-1}\gamma} - L^{-1}\gamma$ and $\tau := \lambda^2/[2\gamma(1 - \lambda)]$. Also let $\tilde{\tau} := \tau/[1 + (\lambda^{-1} - 1)\gamma\tau]$. Then (PP-i) with $\lambda_i \equiv \lambda$ and $\tau_i \equiv \tau$ becomes

$$\begin{cases} x^{i+1} := \operatorname{prox}_{\tilde{\tau} G}(\tilde{x}^i - \tilde{\tau} \nabla F(\bar{x}^i)), \\ \bar{x}^{i+1} := x^{i+1} + \lambda(\lambda^{-1} - 1)(x^{i+1} - x^i), \\ \tilde{x}^{i+1} := (\bar{x}^{i+1} + \gamma\tau(\lambda^{-1} - 1)x^{i+1})/(1 + \gamma\tau(\lambda^{-1} - 1)). \end{cases}$$

Both $G_\gamma(x^N; \widehat{x}) + F(x^N) \to G_\gamma(\widehat{x}; \widehat{x}) + F(\widehat{x})$ and $\zeta^N \to \widehat{x}$ at the linear rate $O((1 - \lambda)^N)$.

Demonstration. To derive the claimed algorithm, we divide the first step of (PP-i) by $1 + \gamma\tau(\lambda^{-1} - 1)$. This yields $\tilde{x}^i - \tilde{\tau} \nabla F(\bar{x}^i) \in \tilde{\tau} \partial G(x^{i+1}) + x^{i+1}$; the rest follows from the definition of $\bar{x}^i$ and the proximal map. Observe that $\lambda \in (0,1)$. Lemma 3.5(ii) with $\epsilon = 0$ and $\phi_0 = 1$ thus proves (3.3) and (3.5), and shows that $\phi_N \tau_N \ge \tau/(1-\lambda)^N$ and $\lambda_N^2 \phi_N \ge \lambda^2/(1-\lambda)^N$. As in Example 3.1, we now obtain (3.20) provided $\tau L \le 1$. With our chosen $\tau = \lambda^2/[2\gamma(1-\lambda)]$, this constraints resolves as our assumed upper bound $\lambda \le \sqrt{L^{-2}\gamma^2 + 2L^{-1}\gamma} - L^{-1}\gamma$. With the growth rates above and $\tau = \lambda^2/[2\gamma(1-\lambda)]$, (3.20) after some rearrangements yields

$$
\begin{aligned}
(1-\lambda)\|\zeta^N - \widehat{x}\|^2 + \gamma^{-1}[(\hat{G}_\gamma + F)(x^N) - (\hat{G}_\gamma + F)(\widehat{x})] \\
\le (1-\lambda)^{N+1}\left(\|x^0 - \widehat{x}\|^2 + \gamma^{-1}[(\hat{G}_\gamma + F)(x^0) - (\hat{G}_\gamma + F)(\widehat{x})]\right).
\end{aligned}
$$

Thus the claimed convergence rates hold. □

Remark 3.6. *Douglas–Rachford splitting for the problem* $\min_{x\in X}\ F(x) + G(x)$ *reads*

$$
\begin{cases}
x^{i+1} = \operatorname{prox}_{\gamma F}(v^i),\\
y^{i+1} = \operatorname{prox}_{\gamma G}(2x^{i+1} - v^i),\\
v^{i+1} = v^i + y^{i+1} - x^{i+1}.
\end{cases}
$$

It can be presented in the form (PP) *with* $\tilde{H}_{i+1} = H$ *for* $u = (x, y, v)$ *and*

$$
H(u) := \begin{pmatrix} \tau\partial F(x) + y - v \\ \tau\partial G(y) + v - x \\ x - y \end{pmatrix} \quad \text{and} \quad M_{i+1} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix}.
$$

With

$$
\Gamma_{i+1} := \begin{pmatrix} 0 & I & -I \\ -I & 0 & I \\ I & -I & 0 \end{pmatrix}, \quad \text{taking instead} \quad M_{i+1} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \lambda_{i+1}^{-1} I \end{pmatrix}.
$$

our approach can be used to construct a corrected inertial Douglas–Rachford splitting. We will, however, not pursue this. Instead, in the next section we take the primal-dual proximal splitting as an example of an algorithm with a non-trivial corrector and Γ_{i+1}.

Inertial Douglas–Rachford splitting has previously been studied in [21]. The "corrected" algorithm derived from our approach will be different. Another accelerated approach is considered in [5]. They apply Douglas–Rachford splitting to H *defined in* (1.10) *by writing it in the form* $H(u) = \partial\hat{G}(u) + \Gamma u$ *for the convex function* $\hat{G}(u) := G(x) + F^*(y)$ *and an anti-symmetric operator* Γ, *as we did in Section 1. What this ingenious approach yields is essentially a doubly over-relaxed PDPS.*

4 INERTIAL PRIMAL-DUAL PROXIMAL SPLITTING

We now return to the saddle point problem (1.4). We suppose $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ are (strongly) convex with factors $\gamma, \rho \ge 0$, and $K \in \mathcal{L}(X;Y)$. Recalling (1.10), for some step length, testing, and inertial parameters $\tau_i, \sigma_{i+1}, \phi_i, \psi_{i+1}, \lambda_i, \mu_{i+1} > 0$, we then take $\tilde{H}_{i+1} = W_{i+1}^{-1}H$ as well as

$$
\text{(4.1a)}\qquad H(u) := \begin{pmatrix} \partial G(x) + K^* y \\ \partial F^*(y) - Kx \end{pmatrix}, \qquad W_{i+1} := \begin{pmatrix} \tau_i I & 0 \\ 0 & \sigma_{i+1} I \end{pmatrix}, \qquad Z_{i+1} := \begin{pmatrix} \phi_i I & 0 \\ 0 & \psi_{i+1} I \end{pmatrix},
$$

$$
\text{(4.1b)}\qquad \Gamma_{i+1} := \begin{pmatrix} \gamma\tau_i I & \tau_i K^* \\ -\sigma_{i+1} K & \rho\sigma_{i+1} I \end{pmatrix}, \qquad \Lambda_{i+1} := \begin{pmatrix} \lambda_i I & 0 \\ 0 & \mu_{i+1} I \end{pmatrix}, \qquad \text{and}
$$

$$
\text{(4.1c)}\qquad M_{i+1} = \begin{pmatrix} I & -\mu_{i+1}^{-1}\tau_i K^* \\ -\lambda_i^{-1}\sigma_{i+1}\omega_i K & I \end{pmatrix} \qquad \text{for} \qquad \omega_i := \frac{\lambda_i \phi_i \tau_i}{\lambda_{i+1}\phi_{i+1}\tau_{i+1}}.
$$

We then observe from (2.2) that the corrector

$$
\text{(4.2)}\qquad \check{M}_{i+1} = \Gamma_{i+1}(\Lambda_{i+1}^{-1} - I) = \begin{pmatrix} \gamma\tau_i(\lambda_i^{-1} - 1)I & \tau_i(\mu_{i+1}^{-1} - 1)K^* \\ -\sigma_{i+1}(\lambda_i^{-1} - 1)K & \rho\sigma_{i+1}(\mu_{i+1}^{-1} - 1)I \end{pmatrix}.
$$

We need to satisfy the conditions of Theorem 2.3 for this setup, in particular (2.4) for some $\mathcal{V}_{k+1}(\widehat{u})$ and $\widehat{u} \in H^{-1}(0)$, and show that the estimate (2.6) is useful, in particular that $Z_{i+1}M_{i+1}$ and $\sum_{i=0}^{N-1}\mathcal{V}_{i+1}(\widehat{u})$ are positive, the former grows at a good rate, and that the latter becomes a useful gap functional. In first instance, we intend to develop it into (a multiple of) the Lagrangian (duality) gap

$$
\text{(4.3)}\qquad \mathcal{G}(x, y; \widehat{x}, \widehat{y}) := \bigl(G(x) + \langle \widehat{y}, Kx\rangle - F^*(\widehat{y})\bigr) - \bigl(G(\widehat{x}) + \langle y, K\widehat{x}\rangle - F^*(y)\bigr).
$$

Recalling G_γ and $(F^*)_\rho$ defined in (3.18), we also introduce the strong convexity adjusted gap

$$
\text{(4.4)}\quad \mathcal{G}_{\gamma,\rho}(x, y; \widehat{x}, \widehat{y}) := \bigl(G_\gamma(x; \widehat{x}) + \langle \widehat{y}, Kx\rangle - (F^*)_\rho(\widehat{y}; \widehat{y})\bigr) - \bigl(G_\gamma(\widehat{x}; \widehat{x}) + \langle y, K\widehat{x}\rangle - (F^*)_\rho(y; \widehat{y})\bigr).
$$

Since the problem $\min_x \max_y\ G_\gamma(x; \widehat{x}) + \langle Kx, y\rangle - (F^*)_\rho(y; \widehat{y})$ has the solution $(\widehat{x}, \widehat{y})$, it is clear that $\mathcal{G}_{\gamma,\rho}$ is non-negative, and zero at $(\widehat{x}, \widehat{y})$. Before proving convergence we derive an explicit algorithm from (4.1).

4.1 Algorithm derivation

With the structure (4.1) fixed, we are ready to develop the skeleton of an explicit algorithm out of (PP-I). Since (PP-I) updates

$$
(\bar{x}^i, \bar{y}^i) = \bar{u}^i := u^i + \Lambda_{i+1}(\Lambda_i^{-1} - I)(u^i - u^{i-1}),
$$

we have

$$
\text{(4.5)}\qquad \bar{x}^i = x^i + \lambda_i(\lambda_{i-1}^{-1} - 1)(x^i - x^{i-1}), \quad \text{and} \qquad \bar{y}^i = y^i + \mu_{i+1}(\mu_i^{-1} - 1)(y^i - y^{i-1}).
$$

Using (4.2) and (4.1) we expand (PP-I) as

$$
\begin{cases} 0 \in \tau_i\partial G(x^{i+1}) + \tau_i K^* y^{i+1} + (x^{i+1} - \bar{x}^i) - \mu_{i+1}^{-1}\tau_i K^*(y^{i+1} - \bar{y}^i) \\ \qquad + \gamma\tau_i(\lambda_i^{-1} - 1)(x^{i+1} - x^i) + \tau_i(\mu_{i+1}^{-1} - 1)K^*(y^{i+1} - y^i), \\ 0 \in \sigma_{i+1}\partial F^*(y^{i+1}) - \sigma_{i+1}Kx^{i+1} - \lambda_i^{-1}\sigma_{i+1}\omega_i K(x^{i+1} - \bar{x}^i) + (y^{i+1} - \bar{y}^i) \\ \qquad + \rho\sigma_{i+1}(\mu_{i+1}^{-1} - 1)(y^{i+1} - y^i) - \sigma_{i+1}(\lambda_i^{-1} - 1)K(x^{i+1} - x^i). \end{cases}
$$

The second line in both inclusions comes from the corrector term. Collecting all instances of the same iterate together, this can be simplified as

$$
\begin{cases}
0 \in \tau_i \partial G(x^{i+1}) + [1+\gamma\tau_i(\lambda_i^{-1}-1)]x^{i+1} - [\bar{x}^i + \gamma\tau_i(1-\lambda_i)x^i] \\
\qquad + \tau_i K^*[\mu_{i+1}^{-1}\bar{y}^i - (\mu_{i+1}^{-1}-1)y^i], \\
0 \in \sigma_{i+1}\partial F^*(y^{i+1}) + [1+\rho\sigma_{i+1}(\mu_{i+1}^{-1}-1)]y^{i+1} - [\bar{y}^i + \rho\sigma_{i+1}(\mu_{i+1}^{-1}-1)y^i] \\
\qquad - \lambda_i^{-1}\sigma_{i+1}(1+\omega_i)Kx^{i+1} + \lambda_i^{-1}\sigma_{i+1}\omega_i K\bar{x}^i + \sigma_{i+1}(\lambda_i^{-1}-1)Kx^i.
\end{cases}
\tag{4.6}
$$

Using (4.5) we can write,

$$
\begin{aligned}
\tilde{y}^i &:= \mu_{i+1}^{-1}\bar{y}^i - (\mu_{i+1}^{-1}-1)y^i = \mu_{i+1}^{-1}y^i + (\mu_i^{-1}-1)(y^i - y^{i-1}) - (\mu_{i+1}^{-1}-1)y^i \\
&= y^i + (\mu_i^{-1}-1)(y^i - y^{i-1}).
\end{aligned}
$$

Similarly, defining

$$
\tilde{x}^{i+1} := x^{i+1} + (\lambda_i^{-1}-1)(x^{i+1}-x^i),
$$

we can write

$$
\begin{aligned}
x_\omega^{i+1} &:= \lambda_i^{-1}(1+\omega_i)x^{i+1} - \lambda_i^{-1}\omega_i\bar{x}^i - (\lambda_i^{-1}-1)x^i \\
&= [\lambda_i^{-1}x^{i+1} - (\lambda_i^{-1}-1)x^i] + \omega_i[\lambda_i^{-1}x^{i+1} - \lambda_i^{-1}\bar{x}^i] \\
&= \tilde{x}^{i+1} + \omega_i[\lambda_i^{-1}x^{i+1} - (\lambda_i^{-1}-1)x^i + (\lambda_i^{-1}-1)x^i - \lambda_i^{-1}\bar{x}^i] \\
&= \tilde{x}^{i+1} + \omega_i(\tilde{x}^{i+1} - \tilde{x}^i).
\end{aligned}
$$

Also introducing

$$
\tilde{\tau}_i := \tau_i/[1+\gamma\tau_i(\lambda_i^{-1}-1)] \quad \text{and} \quad \tilde{\sigma}_{i+1} := \sigma_{i+1}/[1+\rho\sigma_{i+1}(\mu_{i+1}^{-1}-1)],
$$

we now rewrite (4.6) as

$$
\begin{cases}
0 \in \tau_i\partial G(x^{i+1}) + (\tau_i/\tilde{\tau}_i)x^{i+1} - [\bar{x}^i + \gamma\tau_i(\lambda_i^{-1}-1)x^i] + \tau_i K^*\tilde{y}^i, \\
0 \in \sigma_{i+1}\partial F^*(y^{i+1}) + (\sigma_{i+1}/\tilde{\sigma}_{i+1})y^{i+1} - [\bar{y}^i + \rho\sigma_{i+1}(\mu_{i+1}^{-1}-1)y^i] - \sigma_{i+1}Kx_\omega^{i+1}.
\end{cases}
$$

Multiplying, respectively, by $\tilde{\tau}_i/\tau_i$ and $\tilde{\sigma}_{i+1}/\sigma_{i+1}$, and recalling (4.5), we obtain the proximal updates of Algorithm 4.1, which we have written somewhat more compactly by additionally introducing the iterates x_γ^i and y_ρ^i. The updates of $\bar{x}^{i+1}$, $\tilde{x}^{i+1}$, $\bar{y}^{i+1}$, and $\tilde{y}^{i+1}$ in the main step of the method are simply the definitions from above. The step length parameters will still need to be determined from one of the theorems referenced in Algorithm 4.1. Observe how the "corrected" inertial variables $\tilde{x}^{i+1}$ and $\tilde{y}^{i+1}$ differ from the standard inertial variables $\bar{x}^{i+1}$ and $\bar{y}^{i+1}$.

Before developing specific rules for the step lengths and inertial parameters, we still need to provide the estimate (2.4). This process will produce additional conditions on the parameters.

4.2 Basic conditions

We now verify the basic conditions of Theorem 2.3 and the positivity of $Z_{i+1}M_{i+1}$.

Algorithm 4.1. Inertial, corrected, primal-dual proximal splitting (IC-PDPS)

Require: On Hilbert spaces X and Y, a linear operator $K \in \mathcal{L}(X;Y)$ and convex, proper, and lower semicontinuous $G : X \to \overline{\mathbb{R}}$ and $F^* : Y \to \overline{\mathbb{R}}$ with factors $\gamma, \rho \geq 0$ of (strong) convexity.

Determine step length and inertial parameters $\{(\tau_i, \sigma_{i+1}, \lambda_i, \mu_{i+1}, \omega_i)\}_{i\in\mathbb{N}}$ from a suitable one among Theorems 4.5 to 4.8.

Pick initial iterates $\tilde{x}^0 := \bar{x}^0 := x^0 \in X$, and $\tilde{y}^0 := \bar{y}^0 := y^0 \in \operatorname{dom} \partial F^*$.

Let $i := 0$.

repeat

Let

$$\begin{cases} \tilde{\tau}_i := \tau_i/[1 + \gamma\tau_i(\lambda_i^{-1} - 1)], \\ \tilde{\sigma}_{i+1} := \sigma_{i+1}/[1 + \rho\sigma_{i+1}(\mu_{i+1}^{-1} - 1)]. \end{cases}$$

Compute

$$\begin{cases} x_\gamma^i := [\bar{x}^i + \gamma\tau_i(\lambda_i^{-1} - 1)x^i]/[1 + \gamma\tau_i(\lambda_i^{-1} - 1)], \\ x^{i+1} := \operatorname{prox}_{\tilde{\tau}_i G}(x_\gamma^i - \tilde{\tau}_i K^* \tilde{y}^i), \\ \bar{x}^{i+1} := x^{i+1} + \lambda_{i+1}(\lambda_i^{-1} - 1)(x^{i+1} - x^i), \\ \tilde{x}^{i+1} := x^{i+1} + (\lambda_i^{-1} - 1)(x^{i+1} - x^i), \\ y_\rho^i := [\bar{y}^i + \rho\sigma_{i+1}(\mu_{i+1}^{-1} - 1)y^i]/[1 + \rho\sigma_{i+1}(\mu_{i+1}^{-1} - 1)], \\ y^{i+1} := \operatorname{prox}_{\tilde{\sigma}_{i+1} F^*}(y_\rho^i + \sigma_{i+1} K[\tilde{x}^{i+1} + \omega_i(\tilde{x}^{i+1} - \tilde{x}^i)]), \\ \bar{y}^{i+1} := y^{i+1} + \mu_{i+2}(\mu_{i+1}^{-1} - 1)(y^{i+1} - y^i), \\ \tilde{y}^{i+1} := y^{i+1} + (\mu_{i+1}^{-1} - 1)(y^{i+1} - y^i). \end{cases}$$

Update $i := i + 1$

until a stopping criterion is satisfied.

Observe: If $\gamma = 0$, then $\tilde{\tau}_i = \tau_i$ and $x_\gamma^i = \bar{x}^i$. If $\rho = 0$, then $\tilde{\sigma}_{i+1} = \sigma_{i+1}$ and $y_\rho^i = \bar{y}^i$.

Lemma 4.1. *With the setup (4.1), the condition (2.5) holds if*

$$\lambda_i^2 \phi_i (1 + 2\gamma\tau_i\lambda_i^{-1}) \geq \lambda_{i+1}^2 \phi_{i+1}, \tag{4.7a}$$

$$\mu_{i+1}^2 \psi_{i+1} (1 + 2\rho\sigma_{i+1}\mu_{i+1}^{-1}) \geq \mu_{i+2}^2 \psi_{i+2}, \tag{4.7b}$$

$$\lambda_i \phi_i \tau_i = \mu_i \psi_i \sigma_i. \tag{4.7c}$$

Proof. Inserting the operators from (4.1), the condition (2.5) reads

$$\begin{pmatrix} \phi_i\lambda_i(\lambda_i + 2\gamma\tau_i)I & \phi_i\lambda_i\tau_i K^* \\ -\psi_{i+1}\mu_{i+1}\sigma_{i+1}(2 + \omega_i)K & \psi_{i+1}\mu_{i+1}(\mu_{i+1} + 2\rho\sigma_{i+1})I \end{pmatrix} \geq \begin{pmatrix} \phi_{i+1}\lambda_{i+1}^2 I & -\phi_{i+1}\lambda_{i+1}\tau_{i+1}K^* \\ -\psi_{i+2}\mu_{i+2}\sigma_{i+2}\omega_{i+1}K & \psi_{i+2}\mu_{i+2}^2 I \end{pmatrix}.$$

Further inserting ω_i and ω_{i+1} from (4.1c), and using (4.7c), the off-diagonal components cancel out. The diagonal components that are left are simply (4.7a) and (4.7b). □

Lemma 4.2. *Let $i \in \mathbb{N}$. If (4.1) and (4.7) hold, then $Z_{i+1}M_{i+1}$ is self-adjoint. If, moreover,*

$$(1 - \kappa)\mu_{i+1}^2 \psi_{i+1} \geq \phi_i \tau_i^2 \|K\|^2 \quad \textit{for some} \quad \kappa \in [0, 1), \tag{4.8}$$

then $Z_{i+1}M_{i+1}$ is positive definite, more precisely

$$Z_{i+1}M_{i+1} \geq \delta Z_{i+1} \quad \textit{for} \quad \delta := 1 - \sqrt{1-\kappa}. \tag{4.9}$$

Proof. For now, take arbitrary $\delta \in [0, \kappa]$. From (4.1c), using Cauchy's inequality

$$Z_{i+1}M_{i+1} = \begin{pmatrix} \phi_i I & -\mu_{i+1}^{-1}\phi_i\tau_i K^* \\ -\mu_{i+1}^{-1}\phi_i\tau_i K & \psi_{i+1} I \end{pmatrix} \geq \begin{pmatrix} \delta\phi_i I & 0 \\ 0 & \psi_{i+1}I - (1-\delta)^{-1}\mu_{i+1}^{-2}\phi_i\tau_i^2 KK^* \end{pmatrix}.$$

Clearly then $Z_{i+1}M_{i+1}$ is self-adjoint. Using (4.8), we have

$$\psi_{i+1}I - (1-\delta)^{-1}\mu_{i+1}^{-2}\phi_i\tau_i^2 KK^* \geq \psi_{i+1}I - (1-\delta)^{-1}(1-\kappa)\psi_{i+1}I = (\kappa-\delta)(1-\delta)^{-1}\psi_{i+1}I.$$

To make a specific choice of δ, we equate $\delta = (\kappa-\delta)(1-\delta)^{-1}$. This gives the quadratic equation $2\delta - \delta^2 - \kappa = 0$ with the solution $\delta = 1 - \sqrt{1-\kappa}$. The rest is trivial. □

4.3 Gap unrolling and alignment

We now derive a basic convergence estimate using Theorem 2.3. This involves verifying (2.4) for some $\mathcal{V}_{i+1}(\widehat{u})$, and estimating the sum of the latter from below to yield a useful gap estimate. For the statement of the next lemma, we recall the definition of the strong-convexity adjusted functions G_γ and $(F^*)_\rho$ from Section 3.4, and the corresponding gap functional defined in (4.4).

Lemma 4.3. *Suppose (4.1) and (4.7) hold. Take*

$$(1-\lambda_{i+1})\phi_{i+1}\tau_{i+1} \leq \phi_i\tau_i, \qquad \lambda_0 = 1, \tag{4.10a}$$

$$(1-\mu_{i+1})\psi_{i+1}\sigma_{i+1} \leq \psi_i\sigma_i \leq \psi_{i+1}\sigma_{i+1}, \qquad \mu_0 = 1, \quad \textit{and} \tag{4.10b}$$

$$\phi_i\tau_i = \psi_i\sigma_i \qquad (i \in \mathbb{N}). \tag{4.10c}$$

Then for any $\widehat{u} = (\widehat{x}, \widehat{y}) \in H^{-1}(0)$, the iterates generated by Algorithm 4.1 satisfy

$$\frac{1}{2}\|z^N - \widehat{u}\|^2_{\Lambda_{N+1}^* Z_{N+1}M_{N+1}\Lambda_{N+1}} + \phi_{N-1}\tau_{N-1}\mathcal{G}_{\gamma,\rho}(u^N; \widehat{u}) \leq C_0(\widehat{u}) \quad (N \geq 1), \tag{4.11}$$

where for any $w^0 \in \partial(F^*)_\rho(y^0)$ we set

$$C_0(\widehat{u}) := \frac{1}{2}\|z^0 - \widehat{u}\|^2_{\Lambda_1^* Z_1 M_1 \Lambda_1} + \psi_0\sigma_0\langle w^0 - K\widehat{x}, y^0 - \widehat{y}\rangle.$$

Proof. Observe that Algorithm 4.1 explicitly requires $y^0 \in \operatorname{dom} \partial F^*$, so some $w^0 \in \partial(F^*)_\rho(y^0; \widehat{y})$ exists. The proximal steps moreover ensure $y^{i+1} \in \operatorname{dom} \partial F^*$ and $x^{i+1} \in \operatorname{dom} \partial G$ for all $i \in \mathbb{N}$.

By the defining (2.3), the auxiliary sequence $\{z^i := (\zeta^i, \eta^i)\}_{i\in\mathbb{N}} \subset X \times Y$ satisfies

$$\lambda_i\zeta^{i+1} = x^{i+1} - (1-\lambda_i)x^i \quad \text{and} \quad \mu_{i+1}\eta^{i+1} = y^{i+1} - (1-\mu_{i+1})y^i \quad (i \in \mathbb{N}). \tag{4.12}$$

Since $\mu_0 = 1$ and $\eta^0 = y^0$, the latter also works for $i = -1$ and any, superfluous, $y^{-1} \in Y$. We observe that

$$\tilde{H}_{i+1}(u^{i+1}) - \Gamma_{i+1}(u^{i+1} - \widehat{u}) = \begin{pmatrix} \tau_i[\partial G(x^{i+1}) - \gamma(x^{i+1} - \widehat{x}) + K^*\widehat{y}] \\ \sigma_{i+1}[\partial F^*(y^{i+1}) - \rho(y^{i+1} - \widehat{y}) - K\widehat{x}] \end{pmatrix}.$$

Let us define (recall Section 3.4)

$$\bar{G}_\gamma(x;\widehat{u}) := G(x) - \frac{\gamma}{2}\|x-\widehat{x}\|^2 + \langle K^*\widehat{y}, x-\widehat{x}\rangle = G_\gamma(x;\widehat{x}) + \langle K^*\widehat{y}, x-\widehat{x}\rangle, \text{ and} \tag{4.13a}$$

$$(\bar{F}^*)_\rho(y;\widehat{u}) := F^*(y) - \frac{\rho}{2}\|y-\widehat{y}\|^2 - \langle K\widehat{x}, y-\widehat{y}\rangle = (F^*)_\rho(y;\widehat{y}) - \langle K\widehat{x}, y-\widehat{y}\rangle. \tag{4.13b}$$

Then $\bar{G}_\gamma$ and $(\bar{F}^*)_\rho$ are convex with $\bar{G}_\gamma(x;\widehat{u}) \ge \bar{G}_\gamma(\widehat{x};\widehat{u})$, and $(\bar{F}^*)_\rho(y;\widehat{u}) \ge (\bar{F}^*)_\rho(\widehat{y};\widehat{u})$ for all $x \in X$ and $y \in Y$. Moreover, (2.4) holds with

$$\begin{aligned}\mathcal{V}_{i+1}(\widehat{u}) &:= \inf \langle \tilde{H}_{i+1}(u^{i+1}) - \Gamma_{i+1}(u^{i+1}-\widehat{u}), z^{i+1}-\widehat{u}\rangle_{\Lambda^*_{i+1}Z_{i+1}} \\ &= \inf \left[\phi_i\tau_i\lambda_i\langle \partial\bar{G}_\gamma(x^{i+1};\widehat{u}), \zeta^{i+1}-\widehat{x}\rangle + \psi_{i+1}\sigma_{i+1}\mu_{i+1}\langle \partial(\bar{F}^*)_\rho(y^{i+1};\widehat{u}), \eta^{i+1}-\widehat{y}\rangle\right].\end{aligned}$$

For each $i \in \mathbb{N}$, let $\bar{q}^{i+1} \in \partial\bar{G}_\gamma(x^{i+1};\widehat{u})$, and let $w^i \in Y$ be such that $\bar{w}^i = w^i - K\widehat{x} \in \partial(\bar{F}^*)_\rho(y^i;\widehat{u})$. Define

$$s_N := s_N^G + s_N^{F^*} := \sum_{i=0}^{N-1} \phi_i\tau_i\lambda_i\langle \bar{q}^{i+1}, \zeta^{i+1}-\widehat{x}\rangle + \sum_{i=0}^{N-1} \psi_{i+1}\sigma_{i+1}\mu_{i+1}\langle \bar{w}^{i+1}, \eta^{i+1}-\widehat{y}\rangle. \tag{4.14}$$

Since we have assumed (4.1) and (4.7), we may use Lemmas 4.1 and 4.2 to verify (2.5) and the self-adjointness of $Z_{i+1}M_{i+1}$. We may therefore use Theorem 2.3 to establish (4.11) if we further show that

$$s_N \ge \phi_{N-1}\tau_{N-1}\mathcal{G}_{\gamma,\rho}(u^N;\widehat{u}) - c_0 \tag{4.15}$$

for

$$c_0 := \psi_0\sigma_0\mu_0\langle \bar{w}^0, \eta^0-\widehat{y}\rangle = \psi_0\sigma_0\langle w^0 - K\widehat{x}, y^0-\widehat{y}\rangle.$$

Indeed, this establishes the right hand side as a lower bound on $\sum_{i=0}^{N-1}\mathcal{V}_{i+1}(\widehat{u})$.

The difficulty in working with s_N is that unless $\gamma=\rho=0$, our algorithm will give $\phi_i\tau_i = \psi_i\sigma_i$, not $\phi_i\tau_i = \psi_{i+1}\sigma_{i+1}$. We therefore have to realign variables. Using the assumption $\lambda_0 = 1$, (4.10a), and (4.12), by Lemma 3.2 and (4.13a) we have

$$\begin{aligned} s_N^G &\ge \phi_{N-1}\tau_{N-1}[\bar{G}_\gamma(x^N;\widehat{u}) - \bar{G}_\gamma(\widehat{x};\widehat{u})] \\ &= \phi_{N-1}\tau_{N-1}[G_\gamma(x^N;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] + \phi_{N-1}\tau_{N-1}\langle K^*\widehat{y}, x^N-\widehat{x}\rangle.\end{aligned} \tag{4.16}$$

Regarding $s_N^{F^*}$, by the second inequality of (4.10b) we have $\psi_N\sigma_N \ge \psi_{N-1}\sigma_{N-1}$, Moreover, $\widehat{y}$ minimises $(\bar{F}^*)_\rho(\,\cdot\,;\widehat{u})$. Using these two facts after an application analogous to (4.16) of Lemma 3.2, we get

$$\begin{aligned} s_N^{F^*} &= \sum_{i=0}^{N} \psi_i\sigma_i\mu_i\langle \bar{w}^i, \eta^i-\widehat{y}\rangle - c_0 \\ &\ge \psi_N\sigma_N[(\bar{F}^*)_\rho(y^N;\widehat{u}) - (\bar{F}^*)_\rho(\widehat{y};\widehat{u})] - c_0 \\ &\ge \psi_{N-1}\sigma_{N-1}[(\bar{F}^*)_\rho(y^N;\widehat{u}) - (\bar{F}^*)_\rho(\widehat{y};\widehat{u})] - c_0 \\ &= \psi_{N-1}\sigma_{N-1}[(F^*)_\rho(y^N;\widehat{y}) - (F^*)_\rho(\widehat{y};\widehat{y})] - \psi_{N-1}\sigma_{N-1}\langle K\widehat{x}, y^N-\widehat{y}\rangle - c_0.\end{aligned} \tag{4.17}$$

Combining (4.14), (4.16), and (4.17), thus

$$\begin{aligned} s_N \ge{}& \phi_{N-1}\tau_{N-1}[G_\gamma(x^N;\widehat{x}) - G_\gamma(\widehat{x};\widehat{x})] + \psi_{N-1}\sigma_{N-1}((F^*)_\rho(y^N;\widehat{y}) - (F^*)_\rho(\widehat{y};\widehat{y})) \\ &+ \phi_{N-1}\tau_{N-1}\langle K^*\widehat{y}, x^N-\widehat{x}\rangle - \psi_{N-1}\sigma_{N-1}\langle K\widehat{x}, y^N-\widehat{y}\rangle - c_0.\end{aligned}$$

Now (4.10c) establishes (4.15). ◻

4.4 STEP LENGTH AND INERTIAL PARAMETER RULES

We now consider several cases of the factors of (strong) convexity ρ and γ being zero or positive. Throughout, as in the proof of (4.3), we write $z^i = (\zeta^i, \eta^i)$, for the auxiliary sequence $\{z^i\}_{i\in\mathbb{N}}$ defined in (2.3). We first summarise the various lemmas and their conditions from above.

Lemma 4.4. *With $\lambda_0 = 1$ and $\tau_0, \sigma_0, \phi_0, \psi_0 > 0$, suppose that $\mu_i = \lambda_i$ as well as*

$$\psi_i\sigma_i = \phi_i\tau_i, \qquad \omega_i\lambda_{i+1}\phi_{i+1}\tau_{i+1} = \lambda_i\phi_i\tau_i. \tag{4.18a}$$

$$\lambda_i^2\phi_i(1 + 2\gamma\tau_i\lambda_i^{-1}) \geq \lambda_{i+1}^2\phi_{i+1}, \qquad (1 - \lambda_{i+1})\phi_{i+1}\tau_{i+1} \leq \phi_i\tau_i \leq \phi_{i+1}\tau_{i+1}, \tag{4.18b}$$

$$\lambda_i^2\psi_i(1 + 2\rho\sigma_i\lambda_i^{-1}) \geq \lambda_{i+1}^2\psi_{i+1}, \qquad (1 - \kappa)\lambda_{i+1}^2\psi_{i+1} \geq \phi_i\tau_i^2\|K\|^2 \qquad (i \in \mathbb{N}) \tag{4.18c}$$

for some $\kappa \in [0, 1)$. Then the iterates generated by Algorithm 4.1 and the auxiliary sequence generated by (2.3) satisfy with $\delta := 1 - \sqrt{1-\kappa}$ for any $N \geq 1$ and any $\widehat{u} \in H^{-1}(0)$ the estimate

$$\frac{\delta\phi_N\lambda_N^2}{2}\|\zeta^N - \widehat{x}\|^2 + \frac{\delta\psi_{N+1}\lambda_{N+1}^2}{2}\|\eta^N - \widehat{y}\|^2 + \phi_{N-1}\tau_{N-1}\mathcal{G}_{\gamma,\rho}(u^N; \widehat{u}) \leq C_0(\widehat{u}). \tag{4.19}$$

Proof. We first show that the setup (4.1) and the conditions (4.7), (4.8), and (4.10) hold. Indeed, the second part of (4.18a) is simply the choice of ω_i in (4.1c), while the rest of (4.1c) follows from the derivation of Algorithm 4.1 from this structural setup in Section 4.1. Moreover, since $\mu_i = \lambda_i$, the first part of (4.18a) implies (4.7c) and (4.10c). Likewise, (4.18b) implies (4.7a) and (4.10a). The conditions (4.18c) in turn imply (4.7b) and (4.8). Together (4.18a) and (4.18b) imply (4.10b). Therefore (4.7), (4.8), and (4.10) hold in their entirety. We can thus apply Lemmas 4.1 and 4.3 to obtain the estimate (4.11). By application of Lemma 4.2 we then derive (4.19) from (4.11). ◻

Theorem 4.5. *Suppose $\gamma = 0$ and $\rho = 0$. Take $\tau_0, \sigma_0 > 0$ with $\tau_0\sigma_0\|K\|^2 < 1$, $\lambda_0 = \mu_0 = 1$, $\epsilon \in [0, 1)$, and update ($i \in \mathbb{N}$)*

$$\tau_{i+1} := \tau_i\lambda_i^{-1}\lambda_{i+1}, \qquad \omega_i := 1, \tag{4.20a}$$

$$\sigma_{i+1} := \sigma_i\lambda_i^{-1}\lambda_{i+1}, \qquad \mu_{i+1} := \lambda_{i+1} := \lambda_i/(1 + (1-\epsilon)\lambda_i). \tag{4.20b}$$

Then the iterates generated by Algorithm 4.1 satisfy $\mathcal{G}(u^N; \widehat{u}) \to 0$ at the rate $O(1/N)$ for any $\widehat{u} \in H^{-1}(0)$.

Proof. We will use Lemma 4.4, for which we need to verify (4.18). We use Lemma 3.5 (iii) to verify (4.18b) for $\phi_i = \tau_i^{-2}$, $\tau_{i+1} = \tau_i(1-\lambda_{i+1})/(1-\epsilon\lambda_i)$, and λ_{i+1} as in (3.14). With $\gamma = 0$, the latter agrees with the expression for λ_{i+1} in (4.20). With ϕ_i and $\rho = 0$ inserted, the rest of (4.18) reads

$$\psi_i\sigma_i = \tau_i^{-1}, \qquad \omega_i\lambda_{i+1}\tau_{i+1}^{-1} = \lambda_i\tau_i^{-1}. \tag{4.21a}$$

$$\lambda_i^2\psi_i \geq \lambda_{i+1}^2\psi_{i+1}, \qquad (1-\kappa)\lambda_{i+1}^2\psi_{i+1} \geq \|K\|^2 \qquad (i \in \mathbb{N}). \tag{4.21b}$$

Clearly the first part of (4.21b) holds by taking $\psi_i = \lambda_i^{-2}\tau_0^{-1}\sigma_0^{-1}$ for all $i \in \mathbb{N}$. Let us assume

$$\omega_i = (\lambda_{i+1}^{-1} - 1)/(\lambda_i^{-1} - \epsilon), \quad \tau_{i+1} = \tau_i\lambda_i^{-1}\lambda_{i+1}\omega_i, \quad \text{and} \quad \sigma_{i+1} := \sigma_i\lambda_i^{-1}\lambda_{i+1}/\omega_i. \tag{4.22}$$

Inserting $\tau_{i+1} = \tau_i(1-\lambda_{i+1})/(1-\epsilon\lambda_i)$ from above and ω_i from (4.22), the second part of (4.21a) holds. It therefore only remains to secure the first part of (4.21a) and the second part of (4.21b). With ψ_i inserted, this is to say

$$\sigma_i\tau_i = \sigma_0\tau_0\lambda_i^2 \quad \text{and} \quad (1-\kappa) \geq \tau_0\sigma_0\|K\|^2.$$

The second condition is simply our initial condition on the step lengths. The first condition holds if $\sigma_i = \tau_i^{-1}\lambda_i^2\sigma_0\tau_0$. Using that $\tau_{i+1} = \tau_i\lambda_i^{-1}\lambda_{i+1}\omega_i$, this holds when σ_{i+1} as in (4.22). We have therefore proved (4.18) to hold when (4.22) does, $\phi_i = \tau_i^{-2}$, $\psi_i = \lambda_i^{-2}\tau_0^{-1}\sigma_0^{-1}$, and $\tau_0\sigma_0\|K\|^2 < 1$.

Take now as stated $\omega_i = 1$, and observe that the update rule for λ_{i+1} in (4.20) gives the rule for ω_i in (4.22). Moreover, the rules for τ_{i+1} and σ_{i+1} in (4.20) are consistent with (4.22). Therefore (4.22), consequently (4.18), holds under the conditions of the theorem and the choices of the testing parameters ϕ_i and ψ_i in the previous paragraph. Lemma 4.4 thus yields (4.19). Since $\mathcal{G}_{\gamma,\rho}(u^N;\widehat{u}) \geq 0$ when $\widehat{u} \in H^{-1}(0)$, the growth estimate of Lemma 3.5 (iii) in the case $\gamma = 0$ applied in (4.19) establishes the claimed convergence rate. □

Thus, without any strong convexity, inertia and correction improve the ergodic $O(1/N)$ convergence of the gap for the PDPS to non-ergodic convergence.

Theorem 4.6. *Suppose $\gamma > 0$ and $\rho = 0$. Take $\epsilon \in [0,1)$, $\tau_0, \sigma_0 > 0$ with $\tau_0\sigma_0\|K\|^2 < 1$, initialise $\lambda_0 := \mu_0 := 1$, and update ($i \in \mathbb{N}$)*

$$\tau_{i+1} := \tau_i\lambda_i^{-1}\lambda_{i+1}\omega_i, \qquad \omega_i := (\lambda_{i+1}^{-1}-1)/(\lambda_i^{-1}-\epsilon), \tag{4.23a}$$

$$\sigma_{i+1} := \sigma_i\lambda_i^{-1}\lambda_{i+1}/\omega_i, \qquad \mu_{i+1} := \lambda_{i+1} := \frac{\sqrt{\lambda_i^2 + 2\gamma\lambda_i\tau_i}}{1-\epsilon\lambda_i + \sqrt{\lambda_i^2+2\gamma\lambda_i\tau_i}}. \tag{4.23b}$$

Then the iterates generated by Algorithm 4.1 satisfy both $\mathcal{G}_{\gamma,0}(u^N;\widehat{u}) \to 0$ and $\|\zeta^N - \widehat{x}\|^2 \to 0$ at the rate $O(1/N^2)$ for any $\widehat{u} \in H^{-1}(0)$.

Proof. Note that λ_{i+1} given in (3.14) agrees with that in (4.23). Also note that in the proof of Theorem 4.5, we did not involve the choices (4.20) until the final paragraph. Therefore, we may follow the proof of Theorem 4.5 to see (4.18) to hold when (4.22) does, $\phi_i = \tau_i^{-2}$, $\psi_i = \lambda_i^{-2}\tau_0^{-1}\sigma_0^{-1}$, and $\tau_0\sigma_0\|K\|^2 < 1$.

Observe now that (4.22) is consistent with the updates of ω_i, τ_{i+1}, and σ_{i+1} in (4.23). By taking the testing parameters ϕ_i and ψ_i as above, we have therefore verified (4.18), so Lemma 4.4 yields (4.19). The growth estimate of Lemma 3.5 (iii) in the case $\gamma > 0$ applied in (4.19) establishes the claimed convergence rates. □

Theorem 4.7. *Suppose $\gamma = 0$ and $\rho > 0$. Take $\tau_0 > 0$ and $\epsilon \in [0,1/2]$ with $\tau_0\|K\|^2 < 2\rho$, initialise $\lambda_0 := \mu_0 := 1$, and update ($i \in \mathbb{N}$)*

$$\tau_i := \tau_0, \qquad \omega_i := (\lambda_{i+1}^{-1}-1)/(\lambda_i^{-1}-\epsilon) = \lambda_{i+1}\lambda_i^{-1},$$

$$\sigma_{i+1} := \frac{\lambda_i^2}{2\rho} \qquad \mu_{i+1} := \lambda_{i+1} := \frac{2}{1+\sqrt{1+4(\lambda_i^{-2}-\epsilon\lambda_i^{-1})}},$$

Then the iterates generated by Algorithm 4.1 satisfy both $\mathcal{G}_{0,\rho}(u^N;\widehat{u}) \to 0$ and $\|\eta^N - \widehat{y}\|^2 \to 0$ at the rate $O(1/N^2)$ for any $\widehat{u} \in H^{-1}(0)$.

Proof. We use Lemma 3.5 (i) to verify (4.18b) for $\phi_i = \lambda_i^{-2}$ as well as $\tau_i = \tau_0$ and λ_{i+1} as stated. Inserting the ϕ_i and τ_i, the rest of (4.18) now reduces to

$$\begin{aligned} \psi_i\sigma_i &= \lambda_i^{-2}\tau_0, & & \omega_i\lambda_{i+1}^{-1} = \lambda_i^{-1}, \\ \lambda_i^2\psi_i(1+2\rho\sigma_i\lambda_i^{-1}) &\geq \lambda_{i+1}^2\psi_{i+1}, \quad \text{and} & & (1-\kappa)\lambda_{i+1}^2\psi_{i+1} \geq \lambda_i^{-2}\tau_0^2\|K\|^2. \end{aligned}$$

The second condition is one version of our update rule for ω_i. We still need to show that the two versions of the rule are equal. If we take $\psi_i := 2\rho\tau_0\lambda_{i-1}^{-2}\lambda_i^{-2}$ and σ_i as stated, introducing the new variable λ_{-1}, not used in the algorithm, the rest becomes

$$\lambda_{i-1}^{-2} + \lambda_i^{-1} \geq \lambda_i^{-2} \quad \text{and} \quad (1-\kappa)2\rho \geq \tau_0\|K\|^2 \quad (i \in \mathbb{N}).$$

The latter condition is satisfied by our initial step length assumption for some $\kappa \in (0,1)$. Since $\lambda_0 = 1$, the first condition holds for $i = 0$ for any $\lambda_{-1} > 0$. By Lemma 3.4, $\lambda_{i-1}^{-2} - \epsilon\lambda_{i-1}^{-1} = \lambda_i^{-2} - \lambda_i^{-1}$ for $i \in \mathbb{N}$. Therefore the first condition holds, and the two expressions for ω_i are equivalent.

We have thus verified (4.18), so Lemma 4.4 gives the estimate (4.19). The growth estimate of Lemma 3.5 (i) applied there establish the claimed gap convergence rate. The convergence rate of the dual auxiliary variable is determined by the rate of growth of $\lambda_{N+1}^2\psi_{N+1} = 2\rho\tau_0\lambda_N^{-2}$. By the same Lemma 3.5 (i), $\phi_N\tau_N = \lambda_N^{-2}\phi_0$ grows at the rate $\Theta(N^2)$, so we get the claimed $O(1/N^2)$ convergence. □

Theorem 4.8. *Suppose $\gamma > 0$ and $\rho > 0$. Take $\lambda \in (0,1)$ and $\epsilon \in [0,1)$ with $\|K\|^2 < 4\gamma\rho(\lambda^{-1} - \epsilon)(\lambda^{-1} - 1)$. Update ($i \in \mathbb{N}$)*

$$\begin{aligned} \tau_i &:= \lambda^2/[2\gamma(1-\lambda)], & \omega_i &:= (\lambda^{-1}-1)/(\lambda^{-1}-\epsilon), \\ \sigma_i &:= \lambda^2/[2\rho(1-\lambda)], & \mu_{i+1} &:= \lambda_{i+1} := \lambda. \end{aligned}$$

Then the iterates generated by Algorithm 4.1 satisfy both $\mathcal{G}_{\gamma,\rho}(u^N;\widehat{u}) \to 0$ and $\|z^N - \widehat{u}\|^2 \to 0$ at a linear rate for (the unique) $\widehat{u} \in H^{-1}(0)$.

Proof. We use Lemma 3.5 (ii) to verify (4.18b) for $\phi_i = c^i$ for $c := (1-\epsilon\lambda)/(1-\lambda) > 1$ as well as $\tau_i \equiv \tau_0$ and $\lambda_{i+1} \equiv \lambda$ as stated. The rest of (4.18) now reduces to

$$\begin{aligned} \psi_i\sigma_i &= \phi_i\tau_0, & & \omega_i c = 1, \\ \psi_i(1+2\rho\sigma_i\lambda^{-1}) &\geq \psi_{i+1}, \quad \text{and} & & (1-\kappa)\lambda^2\psi_{i+1} \geq \phi_i\tau_0^2\|K\|^2. \end{aligned}$$

Clearly our choice of $\omega_i = 1/c$ satisfies the second condition. Taking $\psi_i = \phi_i\rho/\gamma$ and, as stated, $\sigma_i = \gamma\tau_0/\rho = \lambda^2/[2\rho(1-\lambda)]$, the first condition is also satisfied, while the third condition becomes $\phi_{i+1}(1+2\gamma\tau_0\lambda^{-1}) \geq \phi_i$. As $\lambda_i \equiv \lambda$, this is the first part of (4.18b), which we have already verified. The last condition becomes $(1-\kappa)\lambda^2\rho\gamma^{-1}c \geq \tau_0^2\|K\|^2$, which with c and τ_0 expanded is $4(1-\kappa)\gamma\rho(1-\epsilon\lambda)(1-\lambda) \geq \lambda^2\|K\|^2$. This is secured by our assumed bound on $\|K\|$.

We have thus verified (4.18), so Lemma 4.4 yields the estimate (4.19). The growth estimate of Lemma 3.5 (ii) applied there establish the claimed gap and primal variable convergence rates. Since $\psi_i = \phi_i\rho/\gamma$ and $\mu_i = \lambda_i$, the dual variable converges at the same rate as the primal variable. □

Remark 4.9 (Partial gaps). *Convergence of the Lagrangian gap is weak compared to the true duality gap. Suppose the product set $B_x \times B_y \subset X \times Y$ is bounded and contains some $\widehat{u} \in H^{-1}(0)$. In the literature,* partial gaps *are considered,*

$$0 \leq \mathcal{G}(u; B_x, B_y) := \sup_{\bar{x} \in B_x} \inf_{\bar{y} \in B_y} \mathcal{G}(u; (\bar{x}, \bar{y})).$$

Ergodic partial gap estimates can be derived in the non-inertial setting, see [7], because the fact that $\widehat{u} \in H^{-1}(0)$ is never actually needed in the proofs; $\widehat{u} \in X \times Y$ can be any element. Indeed, even in our work, the main reason we need to assume $\widehat{u}$ to be a solution are the final phases of the unrolling Lemmas 3.2 and 3.3. However, because of this, we cannot derive partial gap estimates.

Remark 4.10 (Forward step). *If we want to solve $\min_{x \in X} G(x) + E(x) + F(Kx)$, where E is convex with ∇E L-Lipschitz, using Lemma 3.3, it is possible to incorporate into Algorithm 4.1 a forward step with respect to E: we change the update of x^{i+1} into*

$$x^{i+1} := \operatorname{prox}_{\tilde{\tau}_i G}(x^i_\gamma - \tilde{\tau}_i[K^* \tilde{y}^i + \nabla E(\bar{x}^i)]).$$

In this case, also minding Lemma 4.2, we have to ensure that $\tau_i L \leq 1 - \sqrt{1-\kappa}$. It is not difficult to verify that the update rule of Lemma 3.5 (iii) satisfies $\lambda_{i+1} \in (\epsilon, 1)\lambda_i$, hence $\tau_{i+1} \leq \tau_i$.

In the proofs of Theorems 4.5 and 4.6 we can take $\kappa \in (0,1)$ such that $\sigma_0\tau_0\|K\|^2 = 1-\kappa$, so we are led to the condition $\sqrt{\sigma_0\tau_0}\|K\| + \tau_0 L \leq 1$, which also ensures, hence replaces, the original bound $\sigma_0\tau_0\|K\|^2 < 1$.

Likewise, in the proof of Theorem 4.7, $\tau_0\|K\|^2 = 2\rho(1-\kappa)$ leads to $\sqrt{\tau_0/(2\rho)}\|K\| + \tau_0 L \leq 1$.

In the proof of Theorem 4.8, $4(1-\kappa)\gamma\rho(\lambda^{-1}-\epsilon)(\lambda^{-1}-1) = \lambda^2\|K\|^2$ similarly leads to the replacement initialisation bound $(4\gamma\rho(\lambda^{-1}-\epsilon)(\lambda^{-1}-1))^{-1/2}\|K\| + \tau_0 L \leq 1$.

5 NUMERICAL EXPERIENCE

We study the performance of the proposed algorithm on three image processing and inverse problems: denoising, sparse Fourier inversion, and Positron Emission Tomography (PET), all with total variation regularisation. We also performed experiments on deblurring, where the results were comparable to denoising. Denoising is the most basic image processing task, while sparse Fourier inversion is used for magnetic resonance image reconstruction; see, e.g., [4, 17]. These two problems are of the form

$$\min_{x \in \mathbb{R}^{n_1 n_2}} \frac{1}{2}\|z - Tx\|_2^2 + \beta\|Dx\|_{2,1}, \tag{5.1}$$

where $n_1 \times n_2$ is the size of the unknown image x in pixels, $z \in \mathbb{R}^m$ is the corrupted data, and $\beta > 0$ a regularisation parameter. The matrix $D \in \mathbb{R}^{2n_1n_2 \times n_1n_2}$ is a discretisation of the gradient operator, and $\|g\|_{2,1} := \sum_{i=1}^{n_1n_2} \sqrt{g_{i,1}^2 + g_{i,2}^2}$ for $g = (g_{\cdot,1}, g_{\cdot,2}) \in \mathbb{R}^{2n_1n_2}$. We take D as forward-differences with Neumann boundary conditions.

The operator $T \in \mathbb{R}^{k \times n_1n_2}$ depends on the problem in question: for denoising, $T = I$ is the identity and for sparse Fourier inversion it is the composition $T = S\mathcal{F}$ with a sub-sampling operator $S \in \mathbb{R}^{k\times n_1 n_2}$ and the discrete Fourier transform $\mathcal{F}$. For denoising $k = n_1 n_2$, while for sparse Fourier reconstruction, $k \ll n_1 n_2$.

To implement variants of the PDPS, we note that (5.1) can in all three cases be written in the saddle point form

$$\min_{x\in\mathbb{R}^{n_1 n_2}} \max_{y\in\mathbb{R}^{2n_1 n_2}} \frac{1}{2}\|z - Tx\|_2^2 + \langle Dx, y\rangle - \delta_{\beta B}(y),$$

where $B = B_{\mathbb{R}^2}^{n_1 n_2}$ for $B_{\mathbb{R}^2}$ the Euclidean unit ball in $\mathbb{R}^2$. Since T is in both cases related to a unitary operator, we can easily compute the proximal map of $G(x) := \frac{1}{2}\|z - Tx\|_2^2$.

The PET problem is slightly different. We take as T a discrete Radon transform, each $[Tx]_j$ being the integral of the image x over a line with angle parameter θ_j and displacement r_j. As the efficient and precise realisation of such an operator in general cases is outside the scope of the present work, in our simplified setting, we consider only the four angles $\theta_j \in \{0°, 45°, 90°, 135°\}$ and displacements r_j such that Tx consists of all row sums, all column sums, and all diagonal sums of x rewritten as a $n_1 \times n_2$ matrix. We also change the first fidelity term in (5.1) to model, instead of Gaussian noise, Poisson noise. Finally, we need to force $x \geq 0$. That is, our problem is

$$\min_{x\in[0,\infty)^{n_1 n_2}} \langle Tx, \mathbb{1}\rangle - \langle b, \log(Tx + c)\rangle + \beta\|Dx\|_{2,1},$$

where $\mathbb{1} := (1, \dots, 1) \in \mathbb{R}^k$, $b \in (0, \infty)^k$ is the measured data, and $c \in (0, \infty)^k$ is a background intensity, assumed known. The logarithm is applied componentwise.

Computing the proximal step with respect to the fidelity term is challenging due to the structure of T. We therefore write also this term as a conjugate, observing that $g_j(z) := z - b_j \log(z + c_j)$ has the conjugate $g_j^*(\phi_j) = -b_j + c_j(1 - \phi_j) + b_j \log(b_j/(1 - \phi_j))$. Introducing the additional upper bound $x \leq 1$, this leads to

$$\min_{x\in\mathbb{R}^{n_1 n_2}} \max_{(\phi,y)\in\mathbb{R}^k\times\mathbb{R}^{2n_1 n_2}} \delta_{[0,1]^{n_1 n_2}}(x) + \langle (Tx, Dx), (\phi, y)\rangle - \left(\delta_{\beta B}(y) + \sum_{j=1}^k g_j^*(\phi_j)\right),$$

Without the additional upper bound, this problem arranged as the prototype problem (1.4) would have $G = \delta_{[0,\infty)^{n_1 n_2}}$, which has the conjugate $G^* = \delta_{(-\infty,0]^{n_1 n_2}}$. Although our algorithms guarantee $x^{i+1} \in [0, \infty)^{n_1 n_2}$, the conjugate will cause the true (non-Lagrangian) duality gap

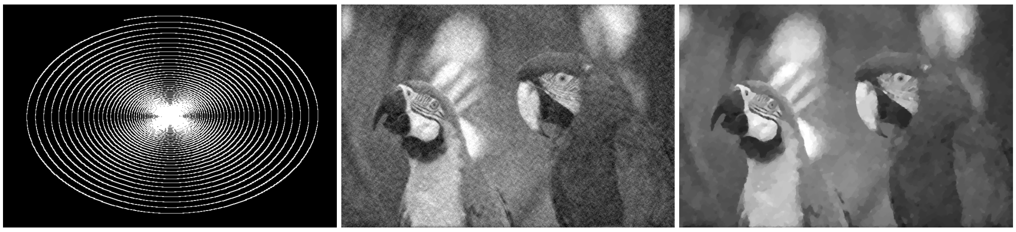
$$\tilde{\mathcal{G}}(x, y) := G(x) + F(Kx) + G^*(-K^* y) + F^*(y) \geq \mathcal{G}(x, y; \widehat{x}, \widehat{y}) \tag{5.2}$$

to be infinite in practise. However, we wish to report the true duality gap instead of the Lagrangian duality gap, as it does not depend on knowing a solution $(\widehat{x}, \widehat{y})$. This is why we have added the upper bound $x \leq 1$. Any greater upper bound would also work, giving a slightly different duality gap.

5.1 DATA

We performed the numerical experiments on the first two of our models on the parrot image (#23) from the free Kodak image suite photo, depicted in Figure 1a together with the corrupted data and restored images for the test problems. We also performed some experiments (see Figure 7)

(a) Original (b) Noisy image (c) Denoised



(d) Fourier subsampling mask (e) Zero-filling inversion (f) Sparse Fourier inversion

Figure 1: Input data and reconstructions. The original image is #23 from the free Kodak image suite, available online at the time of writing at http://rok.us/graphics/kodak/. Since raw data z for the sparse Fourier inversion is not visually informative, (e) displays the naïve zero-filling inversion $\mathcal{F}^*S^*z$ for the subsampling operator S corresponding to the spiral mask in (d).

on all 24 images of this image suite. However, the effect of the exact image on the ranking of the tested algorithms is generally small. The size of all the images is $n_1 \times n_2 = 768 \times 512$. To study scalability, we also scaled it down to $n_1 \times n_2 = 192 \times 128$ pixels. Together with the dual variable, the problem dimensions are therefore $768 \cdot 512 \cdot 3 = 1179648 \simeq 10^6$ and $128 \cdot 128 \cdot 3 = 49152 \approx 4 \cdot 10^4$.

For the denoising problem we added Gaussian noise with standard deviation 51 (−13.9dB) to the original test image. To remove the noise, we first choose $\beta = 0.2$ (low regularisation parameter), and then $\beta = 1$ (high regularisation parameter). Following [12], we scale this parameter by the factor 0.25 for operations on the downscaled image. We also added noise in the other test problems to avoid *inverse crimes* [19]. For sparse Fourier inversion, we used the same level of noise as for denoising. The sparse Fourier inversion experiments are only performed on the original non-down-scaled image with the regularisation parameter $\beta = 0.1$ (sparse Fourier inversion).

For the PET problem, instead of photographs, we use the Shepp–Logan phantom in Figure 2. This is because the limited number of angles encoded in T (reduction of data to mere 2.3% for the phantom) would not give a recognisable reconstruction of a more complex image. Moreover, the phantom is more relevant to the problem in question. As the resolution, we take $n_1 \times n_2 = 256 \times 256$. To obtain the simulated measurement data b, we apply Poisson noise to the row, column and diagonal sums in Tx, and then add the background $c := \mathbb{1}$.

(a) Phantom (b) 4-angle reconstruction

Figure 2: Shepp–Logan brain phantom and its reconstruction from simulated 4-angle (0°, 45°, 90°, 135°) positron emission tomography. The 4-angle tomography of a 256×256 image consists of 1534 data points, meaning the reconstruction is achieved with just 2.3% of data.

5.2 ALGORITHMIC SETUP

We compare our algorithm (IC-PDPS) to the basic PDPS of [7], and the basic inertial (I-PDPS) and over-relaxed (R-PDPS) variants from [9]. The latter is essentially the Vũ–Condat algorithm. We do not include FISTA and other non-primal-dual algorithms in our comparisons, as of our example problems, they are easily applied only to TV denoising in its dual form. Similarly, the basic ADMM [14] requires difficult inversions for our problems. Its more efficient preconditioned variant [27], on the other hand, is equivalent to the PDPS [4].

We use the same initial choices of $\tau_0 = 9.9/L$ and $\sigma_0 = 0.1/L$ with $L := \sqrt{8} \geq \|D\|$ [6] for all algorithms and the denoising and sparse Fourier inversion model problems. For the PET problem we take $\sigma_0 = 30/L'$ and $\tau_0 = 0.033/L'$ for an estimate $L' \geq \sqrt{\|T\|^2 + L^2}$. The ratio between τ_0 and σ_0 has been hand-optimised for the baseline PDPS. For the R-PDPS we take the additional over-relaxation parameter $\rho = 1.5$. For the I-PDPS we use fixed inertial parameter $\alpha = 0.9/3$: according to [9], the sequence of parameters $\{\alpha_i\}_{i\in\mathbb{N}}$ has to be non-decreasing with $\alpha_i < 1/3$. We also tested the FISTA rule, which did in practise yield better results for TV denoising, but completely failed for the other problems. Hence we use the provably convergent fixed parameter.

The denoising problem is strongly convex with factor $\gamma = 1$, so we include results for both the unaccelerated and accelerated versions of the PDPS and IC-PDPS (Theorems 4.5 and 4.6). We also apply the rules of Theorem 4.7 to the problem with the primal and dual variables exchanged. This is denoted 'dual IC-PDPS'. The R-PDPS and the I-PDPS cannot with provable convergence be combined with strong convexity based acceleration: trying to do so was the starting point of our research. For acceleration we use $\gamma = 0.5 < 1$, which is the maximal value for which the ergodic gap is known to converge at the rate $O(1/N^2)$ for the PDPS ($\gamma = 1$ only yields convergence of the iterates; see [7, 22, 23, 25]). For IC-PDPS $\gamma = 1$ is allowed, and provably yields convergence of the gap, but in practise yields worse results than $\gamma = 0.5$.

The IC-PDPS has one further parameter: $\epsilon \in [0, 1)$. For denoising and sparse Fourier inversion we generally take $\epsilon = 0.7$, and for PET, $\epsilon = 0.9$. We also report the denoising convergence behaviour for $\epsilon = 0.5$ and $\epsilon = 0$ in Figure 5.

For our reporting, we computed a target optimal solution $\widehat{x}$ by taking one million iterations of the basic PDPS. However, the convergence of the basic PDPS for sparse Fourier inversion appears to be very slow: judging by the gap in Figure 6a, the IC-PDPS converges much faster, while both the PDPS and I-PDPS flatten out. We therefore computed the target solution for sparse Fourier inversion by taking one million iterations of the IC-PDPS. Note that the target solution is not used to compute the gap; instead of the Lagrangian duality gap (4.3), we report true duality gap given in (5.2), as this does not depend on knowing a solution $(\widehat{x}, \widehat{y})$.

We report the distance to $\widehat{x}$ in decibels $10\log_{10}(\|x^i - \widehat{x}\|^2/\|\widehat{x}\|^2)$, as well as the duality gap, again in decibels relative to the initial gap as $10\log_{10}(\tilde{\mathcal{G}}(x^i, y^i)^2/\tilde{\mathcal{G}}(x^0, y^0)^2)$. For the initial iterates we always took $x^0 = 0$ and $y^0 = 0$. The hardware we used was a MacBook Pro with 16GB RAM and a 2.8 GHz Intel Core i5 CPU. The codes were written in MATLAB+C-MEX.

5.3 RESULTS

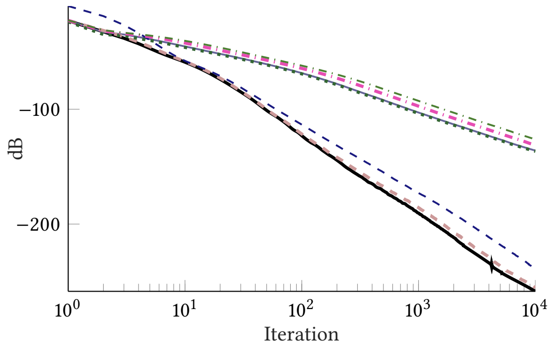
The results for TV denoising of the downscaled image are in Figure 3, and for the original image in Figure 4 and Table 1. The latter includes both the high and low values of the regularisation parameter β. For the downscaled experiments we only report the lower value of β. The comparison for different values of ϵ for IC-PDPS is moreover in Figure 5, for the higher value of β. The results for sparse Fourier inversion are in Figure 6 and Table 4a, and for PET in Figure 8 and Table 4b. Finally, Figure 7 displays for denoising and sparse Fourier inversion the minimum and maximum interval for the duality gap over all 24 images in the image suite. We have excluded R-PDPS from these results to avoid overcrowding; its performance is comparable to I-PDPS, as can be gleaned from the other figures.

For TV denoising, the unaccelerated IC-PDPS is clearly the worst algorithm, while I-PDPS and R-PDPS slightly improve upon the basic PDPS. As expected from the $O(1/N)$ versus $O(1/N^2)$ convergence rates, all of these methods are significantly worse than the accelerated PDPS, the accelerated IC-PDPS, and the accelerated dual IC-PDPS. For the downscaled image and for low β for the original resolution image, they are all comparable for the gap, but accelerated IC-PDPS somewhat surprisingly has asymptotically better iterate convergence. Of course, judging by the timings in Table 1 in particular, the iterations of the IC-PDPS are somewhat more costly, so the basic accelerated PDPS appears the best choice in this case.

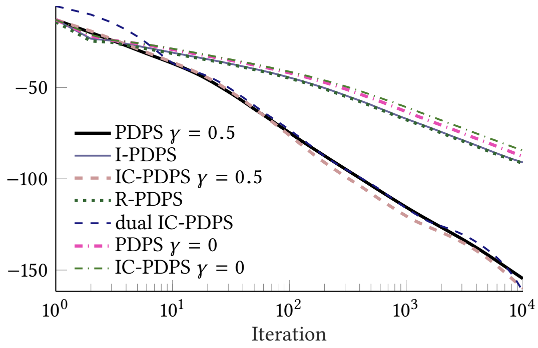
For high β, the results are initially similar, but both variants of the accelerated IC-PDPS are asymptotically better than the accelerated PDPS. This suggests that the IC-PDPS might perform better when there is "more work to be done". This is somewhat confirmed by the results for sparse Fourier inversion, which is a significantly more difficult problem than TV denoising. There the gap convergence performance of IC-PDPS is significantly better than PDPS or I-PDPS: according to Table 4a, compared to the PDPS only 75% of the computational time is required to obtain −35dB gap reduction.

For the PET problem, Figure 8 and Table 4b indicate that IC-PDPS has good gap convergence behaviour, taking 30% less time than the PDPS to reach −40dB, but has primal variable convergence behaviour comparable to the PDPS. This indicates that the IC-PDPS has good convergence of the dual variable.

From Figure 7 we can see that the exact image does not significantly alter the rankings of the



(a) Gap (low resolution, dB)



(b) Distance to target (low resolution, dB)

Figure 3: Denoising convergence behaviour for low resolution image.

algorithms, with IC-PDPS performing significantly better than the other methods for sparse Fourier inversion.

5.4 CONCLUSION

While our proposed IC-PDPS does not always improve upon the basic, inertial, and over-relaxed PDPS, it never does significantly worse by iteration count. For some problems, such as sparse Fourier inversion and Positron Emission Tomography, it offers improved performance. Moreover, we have theoretically guaranteed the $O(1/N)$ convergence of the Lagrangian gap functional or the $O(1/N^2)$ convergence of the strong convexity adjusted gap $\mathcal{G}_{\gamma,\rho}$. This is better than the merely ergodic convergence known of the PDPS and the basic inertial and over-relaxed variants.

ACKNOWLEDGEMENTS

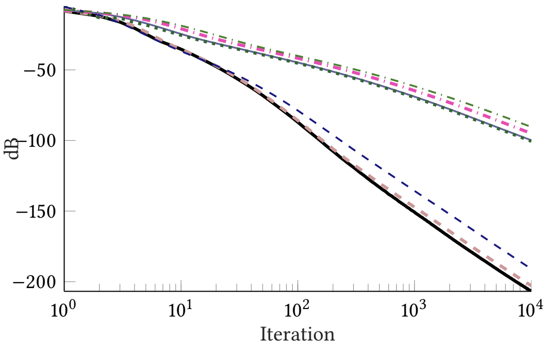
This research has been supported by the EPSRC First Grant EP/P021298/1, "PARTIAL Analysis of Relations in Tasks of Inversion for Algorithmic Leverage" as well as Academy of Finland grants 314701 and 320022.

A DATA STATEMENT FOR THE EPSRC

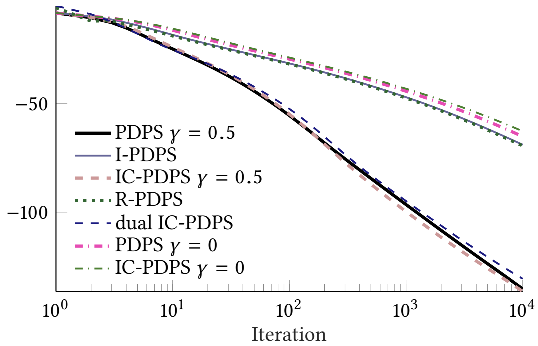
Our algorithm implementations and the publicly available test images have been archived on Zenodo at [24].

REFERENCES

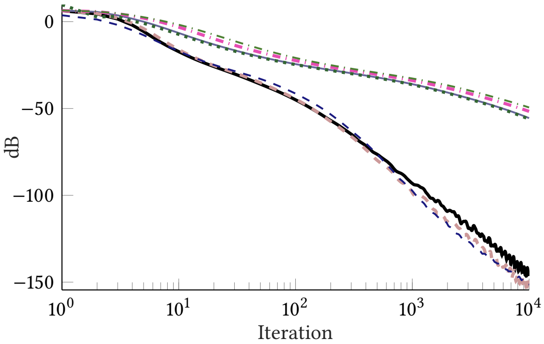
[1] Alvarez & Attouch, An Inertial Proximal Method for Maximal Monotone Operators via Discretization of a Nonlinear Oscillator with Damping, *Set-valued and Variational Analysis* 9 (2001), 3–11, doi: 10.1023/a:1011253113155.



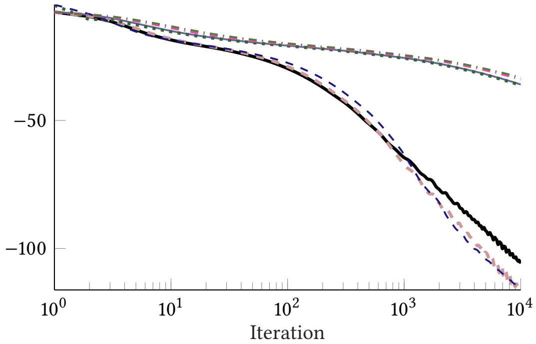
(a) Gap (low β)



(b) Distance to target (low β)

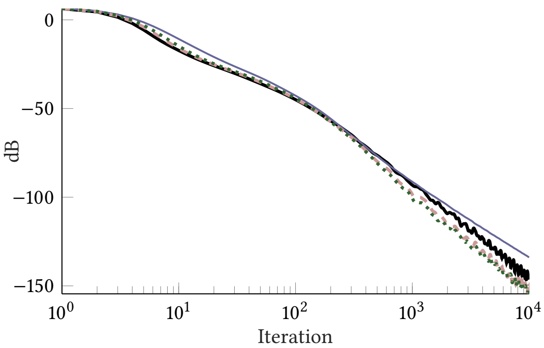


(c) Gap (high β)

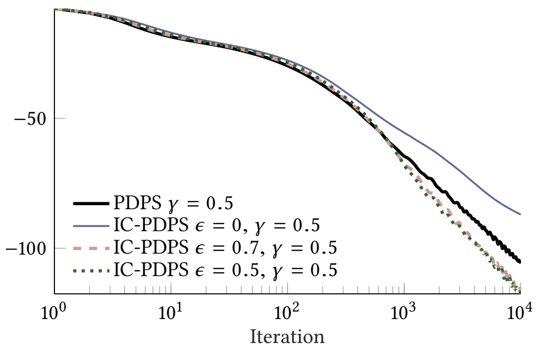


(d) Distance to target (high β)

Figure 4: Denoising convergence behaviour.



(a) Gap

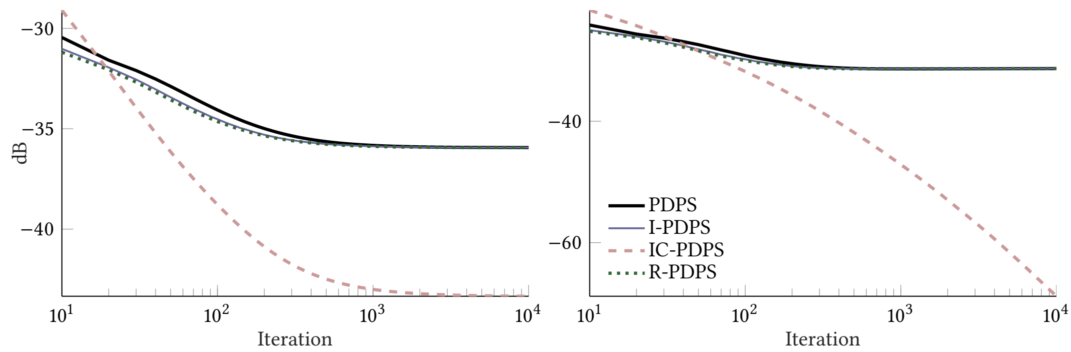


(b) Distance to target

Figure 5: Effect of ϵ on denoising convergence behaviour.

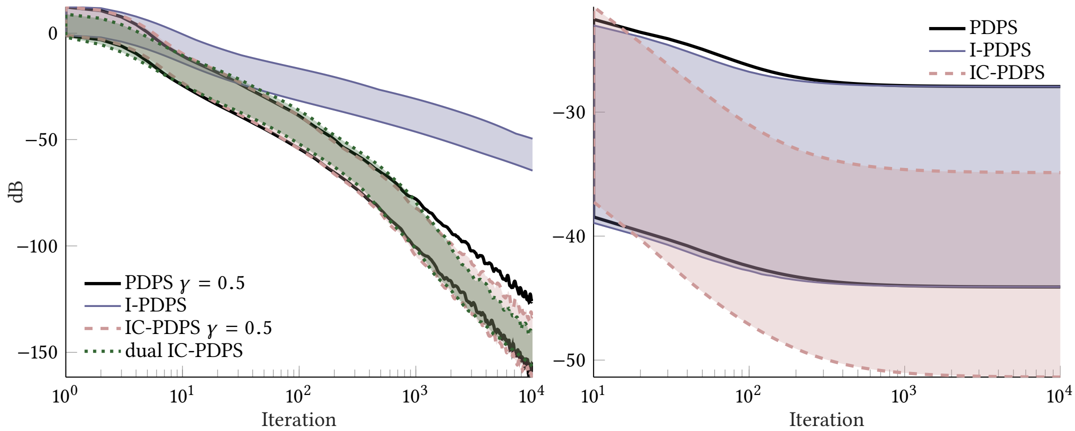
[2] Attouch & Cabot, Convergence Rates of Inertial Forward-Backward Algorithms, *SIAM Journal on Optimization* 28 (2018), 849–874, doi: 10.1137/17m1114739.

[3] Beck & Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *SIAM Journal on Imaging Sciences* 2 (2009), 183–202, doi: 10.1137/080716542.



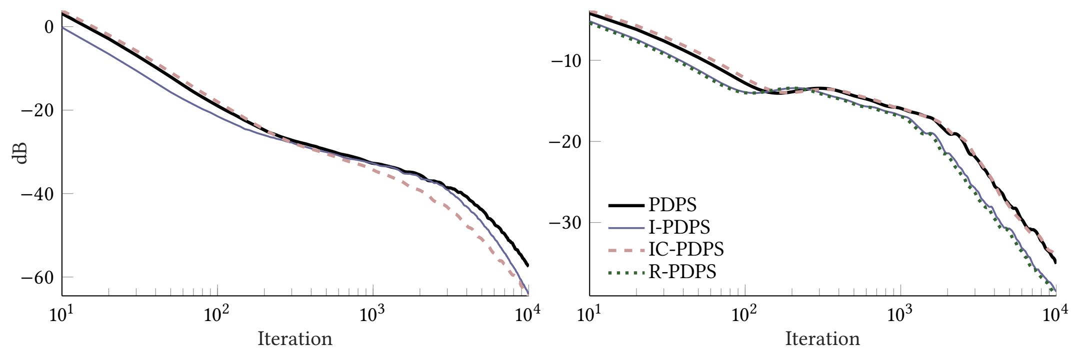
(a) Gap (b) Distance to target*

Figure 6: Sparse Fourier inversion convergence behaviour. *Note: target computed by taking one million iterations of IC-PDPS instead of PDPS; see Section 5.2.



(a) Denoising (b) Sparse Fourier inversion

Figure 7: Gap convergence behaviour over multiple images (24). The filled areas indicate on each iteration the minimum and maximum gap (dB) over all the images.



(a) Gap (b) Distance to target

Figure 8: Convergence behaviour for the PET example problem.

Table 1: Denoising performance: CPU time and number of iterations (at a resolution of 10 after 100 iterations) to reach given duality gap and distance to target. The dashes indicate that the algorithm never reached (within the maximum number of iterations) the corresponding quality.

(a) Low regularisation

	gap ≤ -40dB		gap ≤ -90dB		tgt ≤ -40dB		tgt ≤ -90dB	
Method	iter	time	iter	time	iter	time	iter	time
PDPS $\gamma = 0.5$	14	0.47s	120	4.27s	38	1.33s	690	24.74s
I-PDPS	58	2.39s	4870	203.75s	400	16.70s	–	–
IC-PDPS $\gamma = 0.5$	14	0.75s	120	6.87s	40	2.25s	590	33.99s
R-PDPS	55	2.36s	4630	202.46s	380	16.58s	–	–
dual IC-PDPS	13	0.64s	160	8.53s	43	2.25s	750	40.18s
PDPS $\gamma = 0$	82	2.87s	6950	245.99s	560	19.79s	–	–
IC-PDPS $\gamma = 0$	99	5.00s	9710	494.99s	650	33.09s	–	–

(b) High regularisation

	gap ≤ -40dB		gap ≤ -90dB		tgt ≤ -40dB		tgt ≤ -90dB	
Method	iter	time	iter	time	iter	time	iter	time
PDPS $\gamma = 0.5$	70	2.35s	890	30.34s	250	8.50s	4330	147.73s
I-PDPS	1810	70.74s	–	–	–	–	–	–
IC-PDPS $\gamma = 0.5$	73	3.85s	740	39.51s	270	14.38s	2740	146.43s
R-PDPS	1720	68.07s	–	–	–	–	–	–
dual IC-PDPS	91	4.78s	770	40.85s	330	17.48s	2560	135.94s
PDPS $\gamma = 0$	2580	84.57s	–	–	–	–	–	–
IC-PDPS $\gamma = 0$	3240	157.19s	–	–	–	–	–	–

Table 3: Sparse Fourier inversion and PET performance: CPU time and number of iterations (at a resolution of 10) to reach given duality gap and distance to target. The dashes indicate that the algorithm never reached (within the maximum number of iterations) the corresponding quality.

(a) Sparse Fourier inversion

	gap ≤ -35dB		tgt ≤ -35dB	
Method	iter	time	iter	time
PDPS	210	15.96s	–	–
I-PDPS	150	11.87s	–	–
IC-PDPS	40	3.90s	180	17.88s
R-PDPS	140	12.12s	–	–

(b) PET

	gap ≤ -40dB		tgt ≤ -30dB	
Method	iter	time	iter	time
PDPS	3740	43.31s	6190	71.68s
I-PDPS	3220	48.72s	4380	66.28s
IC-PDPS	2180	30.94s	6010	85.32s
R-PDPS	–	–	4150	57.32s

[4] Benning, Knoll, Schönlieb & Valkonen, Preconditioned ADMM with nonlinear operator constraint, in: *System Modeling and Optimization: 27th IFIP TC 7 Conference, CSMO 2015, Sophia Antipolis, France, June 29–July 3, 2015, Revised Selected Papers*, ed. by Bociu, Désidéri & Habbal, Springer International Publishing, 2016, 117–126, doi: 10.1007/978-3-319-55795-3_10, arxiv: 1511.00425,

[5] Bredies & Sun, Preconditioned Douglas–Rachford Algorithms for TV- and TGV-Regularized Variational Imaging Problems, *Journal of Mathematical Imaging and Vision* 52 (2015), 317–344, doi: 10.1007/s10851-015-0564-1.

[6] Chambolle, An algorithm for total variation minimization and applications, *Journal of Mathematical Imaging and Vision* 20 (2004), 89–97, doi: 10.1023/b:jmiv.0000011325.36760.1e.

[7] Chambolle & 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.

[8] Chambolle & Dossal, On the convergence of the iterates of "FISTA". *Journal of Optimization Theory and Applications* 16 (2015), 25, url: hal.inria.fr/hal-01060130v3.

[9] Chambolle & Pock, On the ergodic convergence rates of a first-order primal–dual algorithm, *Mathematical Programming* (2015), 1–35, doi: 10.1007/s10107-015-0957-3.

[10] Chan, Ma & Yang, Inertial primal-dual algorithms for structured convex optimization, 2014, arxiv: 1409.2992.

[11] Condat, A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms, *Journal of Optimization Theory and Applications* 158 (2013), 460–479, doi: 10.1007/s10957-012-0245-9.

[12] De Los Reyes, Schönlieb & Valkonen, Bilevel parameter learning for higher-order total variation regularisation models, *Journal of Mathematical Imaging and Vision* 57 (2017), 1–25, doi: 10.1007/s10851-016-0662-8, arxiv: 1508.07243,

[13] Eckstein & Bertsekas, On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators, *Mathematical Programming* 55 (1992), 293–318, doi: 10.1007/bf01581204.

[14] Gabay, Applications of the Method of Multipliers to Variational Inequalities, in: *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, ed. by Fortin & Glowinski, North-Holland, 1983, 299–331.

[15] He & 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.

[16] Kim & Fessler, Another Look at the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA), *SIAM Journal on Optimization* 28 (2018), 223–250, doi: 10.1137/16m108940x.

[17] Knoll, Bredies, Pock & Stollberger, Second order total generalized variation (TGV) for MRI, *Magnetic Resonance in Medicine* 65 (2011), 480–491, doi: 10.1002/mrm.22595.

[18] Lorenz & 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.

[19] Mueller & Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications*, SIAM, 2012, doi: 10.1137/1.9781611972344.

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

[21] Patrinos, Stella & Bemporad, Douglas–Rachford splitting: Complexity estimates and accelerated variants, in: *53rd IEEE Conference on Decision and Control*, 2014, 4234–4239, doi: 10.1109/cdc.2014.7040049.

[22] Valkonen, Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics and Optimization* (2018), doi: 10.1007/s00245-018-9541-6, arxiv: 1703.05705,

[23] Valkonen, Block-proximal methods with spatially adapted acceleration, *Electronic Transactions on Numerical Analysis* 51 (2019), 15–49, doi: 10.1553/etna_vol51s15, arxiv: 1609.07373,

[24] Valkonen, Source codes for "Inertial, corrected, primal-dual proximal splitting", Online resource on Zenodo, 2019, doi: 10.5281/zenodo.3531934.

[25] Valkonen & Pock, Acceleration of the PDHGM on partially strongly convex functions, *Journal of Mathematical Imaging and Vision* 59 (2017), 394–414, doi: 10.1007/s10851-016-0692-2, arxiv: 1511.06566,

[26] Vũ, A splitting algorithm for dual monotone inclusions involving cocoercive operators, *Advances in Computational Mathematics* 38 (2013), 667–681, doi: 10.1007/s10444-011-9254-8.

[27] Zhang, Burger & Osher, A unified primal-dual algorithm framework based on Bregman iteration, *Journal of Scientific Computing* 46 (2011), 20–46, doi: 10.1007/s10915-010-9408-8.