ACCELERATION AND GLOBAL CONVERGENCE OF A FIRST-ORDER PRIMAL–DUAL METHOD FOR NONCONVEX PROBLEMS

Christian Clason*　　Stanislav Mazurenko†　　Tuomo Valkonen‡

2018-02-09

Abstract First-order primal–dual algorithms are the backbone for mathematical image processing and more general inverse problems that can be formulated as convex optimization problems. Recent advances have extended their applicability to areas previously dominated by second-order algorithms, such as nonconvex problems arising in optimal control. Nonetheless, the application of first-order primal–dual algorithms to nonconvex large-scale optimization still requires further investigation. In this paper, we analyze an extension of the primal–dual hybrid gradient method (PDHGM, also known as the Chambolle–Pock method) designed to solve problems with a nonlinear operator in the saddle term. Based on the idea of testing, we derive new step length parameter conditions for the convergence in infinite-dimensional Hilbert spaces and provide acceleration rules for suitably locally monotone problems. Importantly, we demonstrate linear convergence rates and prove global convergence in certain cases. We demonstrate the efficacy of these new step length rules on PDE-constrained optimization problems.

1 INTRODUCTION

Many optimization problems can be represented as minimizing a sum of two terms of the form

$$\min_x\ G(x) + F(K(x)). \tag{P}$$

For instance, in inverse problems, G will typically be a fidelity term, measuring fit to data, and $F \circ K$ a regularization term introduced to avoid ill-posedness and promote desired features in the solution. In imaging problems in particular, quite often total variation type regularization is used, in which case K is composed from differential operators [2, 5, 8]. In optimal control, K frequently denotes the solution operator to partial or ordinary differential equations as a function of the control input. In this case G and F stand for control- and state-dependent contributions to the cost function, respectively; the latter might also account for state constraints [12].

Since the applications mentioned above usually involve high and possibly infinite-dimensional spaces, if K can be computed efficiently, first-order numerical methods can provide the best

*Faculty of Mathematics, University Duisburg-Essen, 45117 Essen, Germany (christian.clason@uni-due.de)
†Department of Mathematical Sciences, University of Liverpool, United Kingdom (stan.mazurenko@gmail.com)
‡Department of Mathematical Sciences, University of Liverpool, United Kingdom (tuomo.valkonen@iki.fi)

trade-off between precision and computation time. Moreover, as both G and F are often convex, introducing a dual variable and the convex conjugate F^* of F, we can rewrite (P) as

$$\min_x \max_y \; G(x) + \langle K(x), y\rangle - F^*(y). \tag{S}$$

This approach proves to be particularly successful for nonsmooth G and F^*. Non-smooth first-order methods roughly divide into two classes: ones based on explicit subgradients, and ones based on proximal mappings. The former can exhibit very slow convergence, while taking a step in the latter is often tantamount to solving the original problem. In (S), if we can decouple the primal and dual variables and efficiently compute the proximal maps $(I+\tau\partial G)^{-1}$ and $(I+\sigma\partial F^*)^{-1}$, methods based on proximal maps can become highly efficient. Based on this fact, Chambolle and Pock [7] suggested such a decoupling algorithm for the case that K is linear and proved its convergence to a saddle point with rate $O(1/N)$ in terms of an ergodic primal–dual gap in finite dimensions. They also provided an acceleration scheme with $O(1/N^2)$ rates if the primal or dual objective is strongly convex. In [16], the method was classified as the Primal–Dual Hybrid Gradient method, Modified (PDHGM).

However, frequently in applications, K is not linear. An extension of the PDHGM to nonlinear K was suggested in [12, 21], for which the authors proved a local convergence without a rate under a metric regularity assumption. The method, called the NL-PDHGM for "nonlinear", and its ADMM-form variants, have successfully been applied to problems in magnetic resonance imaging and PDE-constrained optimization [4, 12, 21, 26]. We state NL-PDHGM in Algorithm 1.1 incorporating references to the step length rules of the present work.

Algorithm 1.1 (Exact NL-PDHGM). Pick a starting point (x_0, y_0). Select step length parameters $\tau_i, \sigma_i, \omega_i > 0$ according to suitable rules from Theorems 4.5, 4.7, 4.10, 4.14, 4.16 and 4.18 and Corollaries 4.22 to 4.24. Iterate:

$$\begin{aligned}
x^{i+1} &:= (I + \tau_i\partial G)^{-1}(x^i - \tau_i[\nabla K(x^i)]^* y^i),\\
\bar{x}^{i+1} &:= x^{i+1} + \omega_i(x^{i+1} - x^i),\\
y^{i+1} &:= (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1}K(\bar{x}^{i+1})).
\end{aligned}$$

In [12], based on small modifications to our original analysis in [21], we showed that the $O(1/N^2)$ acceleration scheme from [7] for strongly convex problems can also be used with Algorithm 1.1 if we stop the acceleration at some iteration. At that point, we were unable to provide any convergence rates. In this paper, we provide such rates and show that the acceleration does not have to be stopped. We also present new step length bounds that guarantee convergence, sometimes even globally, and provide criteria for linear convergence.

Our new analysis of the NL-PDHGM is based on the "testing" framework introduced in [24, 25] for preconditioned proximal point methods and summarized in Section 2. In particular, we relax the metric regularity required in [21] to mere monotonicity *at* a solution. We state our main results in Section 3. Since block-coordinate methods have been receiving more and more attention lately – including in the primal–dual algorithm designed in [22] based on the same testing framework – the main technical derivations of Section 3.2 are implemented in a generalized operator form.

Once those generic estimates are obtained, we devote Section 4 to scalar step length parameters and formulate our main convergence results. These amount to basically standard step length rules for the PDHGM combined with bounds on the initial step lengths. We prove weak and strong convergence to a critical point as well as $O(1/N^2)$ convergence with an acceleration rule if ∂G or $[\nabla K(x)]^* y$ is strongly monotone *at* a primal critical point $\widehat{x}$. If ∂F^* is also strongly monotone at a dual critical point $\widehat{y}$, we present step length rules that lead to linear convergence. We then refine the results to the case when $x \mapsto \langle K(x), \widehat{y}\rangle$ has a hypomonotone gradient, e.g., is convex. This connects our work to the classical forward–backward splitting method as well as to the PDHGM with a forward step [9].

Finally, in Section 5, we illustrate our theoretical results with numerical evidence. We study parameter identification with L^1 fitting and optimal control with state constraints, where the nonlinear operator K involves the mapping from a potential term in an elliptic partial differential equation to the corresponding solution.

2 PROBLEM FORMULATION

Throughout this paper, we write $\mathcal{L}(X;Y)$ for the space of bounded linear operators between Hilbert spaces X and Y; I is the identity operator; $\langle x, x'\rangle$ is the inner product in the corresponding space; and $\mathbb{B}(x,r)$ is the closed unit ball of the radius r at x. We set $\langle x, x'\rangle_T := \langle Tx, x'\rangle$, and $\|x\|_T := \sqrt{\langle x, x\rangle_T}$. For $T, S \in \mathcal{L}(X;Y)$, the inequality $T \geq S$ means $T - S$ is positive semidefinite. Finally, $[\![x_1, x_2]\!]^\alpha := (1-\alpha)x_1 + \alpha x_2$, consequently, $\bar{x}^{i+1} := [\![x^{i+1}, x^i]\!]^{-\omega_i}$ in Algorithm 1.1.

We generally assume $G : X \to \overline{\mathbb{R}}$ and $F^* \to \overline{\mathbb{R}}$ to be convex, proper, and lower semicontinuous, so that their subgradients ∂G and ∂F^* would be well-defined maximally monotone operators [3, Theorem 20.25]. We may, therefore, define the set-valued operator $H : X \times Y \rightrightarrows X \times Y$ for $u = (x, y)$ as

$$H(u) := \begin{pmatrix} \partial G(x) + [\nabla K(x)]^* y \\ \partial F^*(y) - K(x) \end{pmatrix}. \tag{2.1}$$

Then $0 \in H(\widehat{u})$ encodes the critical point conditions for (P) and (S). These will also become the first-order necessary optimality conditions under a constraint qualification, e.g., when G is C^1 and either the null space of $[\nabla K(x)]^*$ is trivial or $\operatorname{dom} F = X$ [20, Example 10.8].

To formulate Algorithm 1.1 in terms suitable for the testing framework of [24], we define the step length and testing operators

$$W_{i+1} := \begin{pmatrix} T_i & 0 \\ 0 & \Sigma_{i+1} \end{pmatrix}, \quad \text{and} \quad Z_{i+1} := \begin{pmatrix} \Phi_i & 0 \\ 0 & \Psi_{i+1} \end{pmatrix},$$

where $T_i, \Phi_i \in \mathcal{L}(X;X)$ and $\Sigma_{i+1}, \Psi_{i+1} \in \mathcal{L}(Y;Y)$ are the primal and dual step length and testing operators, respectively.

We also define the nonlinear preconditioner $V_{i+1}(u) := V'_{i+1}(u) + M_{i+1}(u - u^i)$ by

$$V'_{i+1}(u) := W_{i+1} \begin{pmatrix} [\nabla K(x^i) - \nabla K(x)]^* y \\ K(x) - K([\![x, x^i]\!]^{-\omega_i}) - \nabla K(x^i)(x - [\![x, x^i]\!]^{-\omega_i}) \end{pmatrix}, \quad \text{and} \tag{2.2}$$

$$M_{i+1} := \begin{pmatrix} I & -T_i[\nabla K(x^i)]^* \\ -\omega_i \Sigma_{i+1} \nabla K(x^i) & I \end{pmatrix}. \tag{2.3}$$

As we recall, $[\![x, x^i]\!]^{-\omega} = x + \omega(x - x^i)$. Since $V'_{i+1}(u)$ vanishes for linear K, we will also make use of the subspace of Y, possibly empty, in which K acts linearly. In other words, P_{NL} will denote the orthogonal projection to Y_{NL} where

$$Y_{\mathrm{L}} := \{y \in Y \mid \text{the map } x \mapsto \langle y, K(x)\rangle \text{ is linear}\} \quad \text{and} \quad Y_{\mathrm{NL}} := Y_{\mathrm{L}}^{\perp}.$$

See [21] for how such subspaces come about. We also write $\mathbb{B}_{\mathrm{NL}}(\widehat{y}, r) := \{y \in Y \mid \|y - \widehat{y}\|_{P_{\mathrm{NL}}} \leq r\}$ for a closed cylinder in Y of the radius r with axis orthogonal to Y_{NL}.

Now the "exact" NL-PDHGM of [21] can be written as

$$0 \in W_{i+1}H(u^{i+1}) + V_{i+1}(u^{i+1}) =: \widetilde{H}_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - u^i). \tag{PP}$$

For the "linearized" NL-PDHGM of [21], we would replace $[\![x, x^i]\!]^{-\omega}$ in (2.2) by x^i.

In line with [24], the step length operator W_{i+1} in (PP) acts on H rather than on the step $u^{i+1} - u^i$ so as to eventually allow zero-length steps on sub-blocks of variables; cf. [22]. The testing operator Z_{i+1} does not yet appear in (PP): it does not feature in the algorithm. We will shortly see that when we apply it to (PP), the product $Z_{i+1}M_{i+1}$ will form a metric that encodes convergence rates (in the differential-geometric sense of the word "metric").

Accordingly, our goal in the rest of the paper is to analyze the convergence of (PP) for the choices (2.1)–(2.3). We will base this analysis on the following abstract result, which is relatively trivial to prove based on telescoping and Pythagoras' (three-point) formula:

Theorem 2.1 ([24, Theorem 2.1]). *Suppose (PP) is solvable, and denote the iterates by $\{u^i\}_{i\in\mathbb{N}}$. If $Z_{i+1}M_{i+1}$ is self-adjoint, and with $\widetilde{H}_{i+1}$ as in (PP) we have*

$$\frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}} + \langle \widetilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} \geq -\Delta_{i+1} \tag{CI}$$

for all $i \leq N - 1$ and some $\widehat{u} \in U$, then

$$\frac{1}{2}\|u^N - \widehat{u}\|^2_{Z_{N+1}M_{N+1}} \leq \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1M_1} + \sum_{i=0}^{N-1} \Delta_{i+1}. \tag{DI}$$

Clearly, if $\Delta_{i+1} \leq 0$, the rate of convergence is defined by $Z_{N+1}M_{N+1}$ since if $Z_{N+1}M_{N+1} \geq \mu_N I$ and $\mu_N \to \infty$, then $\|u^N - \widehat{u}\|^2 \to 0$ at the rate $O(1/\mu_N)$. If $Z_{N+1}M_{N+1}$ does not grow quickly, we can still obtain weak convergence as follows:

Proposition 2.2 (Weak convergence). *Suppose the iterates of (PP) satisfy (CI) for some $\widehat{u} \in H^{-1}(0)$ with $\Delta_{i+1} \leq -\frac{\hat{\delta}}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$ for some $\hat{\delta} > 0$. If the following conditions hold, then $u^i \rightharpoonup u^*$ weakly in U for some $u^* \in H^{-1}(0)$:*

(i) $\epsilon I \leq Z_{i+1}M_{i+1}$ for some $\epsilon > 0$.

(ii) For some nonsingular $W \in \mathcal{L}(U; U)$ holds

$$Z_{i+1}M_{i+1}(u^{i+1} - u^i) \to 0, \quad u^{i_k} \rightharpoonup u^* \implies 0 \in WH(u^*).$$

(iii) *There exists a constant C such that $\|Z_iM_i\| \le C^2$ for all i, and for any subsequence $u^{i_k} \rightharpoonup u$ there exists $A_\infty \in \mathcal{L}(U;U)$ such that $Z_{i_k+1}M_{i_k+1}u \to A_\infty u$ strongly in U for all $u \in U$.*

Proof. This is an improvement of [24, Proposition 2.5] that permits nonconstant $Z_{i+1}M_{i+1}$ and a nonconvex solution set. The proof is based on the corresponding improvement of Opial's lemma (Lemma A.2) together with Theorem 2.1. Using $\Delta_{i+1} \le -\frac{\hat\delta}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}}$, (DI) applied with $N = 1$ and u^i in place of u^0 shows that $i \mapsto \|u^i - \widehat{u}\|^2_{Z_{i+1}M_{i+1}}$ is nonincreasing. This verifies Lemma A.2(i). Further use of (DI) shows $\sum_{i=0}^\infty \frac{\hat\delta}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} < \infty$. Thus $Z_{i+1}M_{i+1}(u^{i+1} - u^i) \to 0$. By (PP) and (ii), any weak limit point u^* of the $\{u^i\}_{i\in\mathbb{N}}$ therefore satisfies $u^* \in H^{-1}(0)$. This verifies Lemma A.2(ii) with $\hat X = H^{-1}(0)$. The remaining assumptions of Lemma A.2 are verified by conditions (i) and (iii) of the present proposition. Thus, the lemma shows that $u^i \rightharpoonup u^* \in H^{-1}(0)$. □

3 ABSTRACT ANALYSIS OF THE NL-PDHGM

We will apply Theorem 2.1 to Algorithm 1.1, for which we have to verify (CI). Obviously, this inequality holds for some Δ_{i+1}, but we want to make Δ_{i+1} as small as possible. Indeed, we aim for $\Delta_{i+1} \le 0$. To obtain fast convergence rates, our second goal is to make the metric $Z_{i+1}M_{i+1}$ grow as quickly as possible. Since this rate is constrained by the term $\frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}M_{i+1}-Z_{i+2}M_{i+2}}$ in (CI), we deal with this constraint in the present section. The actual convergence rates are, however, only derived in Section 4 for scalar step lengths.

After stating our fundamental assumptions in Section 3.1, we first derive in Section 3.2 explicit – albeit somewhat technical – bounds on the step length operators to ensure (CI). These require that the iterates $\{u^i\}_{i\in\mathbb{N}}$ stay in a neighborhood of the critical point $\widehat{u}$. Therefore, in Section 3.3, we provide sufficient conditions for this requirement to hold in the form of additional step length bounds. These conditions will further be used in Section 4.

3.1 FUNDAMENTAL ASSUMPTIONS

In what follows, we will need to assume that K is Fréchet differentiable and its gradient ∇K is Lipschitz in some neighborhood $\mathcal{X}_K$ of the primal optimal point $\widehat{x}$. Moreover, we assume a form of hypomonotonicity of $x \mapsto \nabla K(x)^*\widehat{y}$, which we will first need later on, in Lemma 3.5.

Assumption 3.1. For some $L \ge 0$, $\Theta \in \mathcal{L}(X;X)$, and a neighborhood $\mathcal{X}_K$ of $\widehat{x}$:

$$\|\nabla K(x) - \nabla K(x')\| \le L\|x - x'\| \qquad (x, x' \in \mathcal{X}_K), \quad \text{and} \tag{3.1a}$$

$$\langle(\nabla K(x) - \nabla K(\widehat{x}))(x - \widehat{x}), \widehat{y}\rangle \ge \|x - \widehat{x}\|^2_\Theta \qquad (x \in \mathcal{X}_K). \tag{3.1b}$$

Remark 3.1. *Using Assumption 3.1 and the equality*

$$K(x') = K(x) + \nabla K(x)(x' - x) + \int_0^1 (\nabla K(x + s(x' - x)) - \nabla K(x))(x' - x)ds,$$

we obtain the following useful inequality for any $x, x' \in \mathcal{X}_K$ and $y \in \operatorname{dom} F^*$:

$$\langle K(x') - K(x) - \nabla K(x)(x' - x), y\rangle \le (L/2)\|x - x'\|^2\|y\|_{P_{\mathrm{NL}}}. \tag{3.2}$$

The norm in the dual space consists of only the Y_{NL} component because by the definition of its complement Y_{L}, the function $x \mapsto \langle K(x), y\rangle$ is linear in x for $y \in Y_{\mathrm{L}}$. Consequently, for such y, the left-hand side of (3.2) *is zero.*

We will also assume a form of monotonicity from ∂G and ∂F^* which we will likewise first need in Lemma 3.5.

Definition 3.2. Let U be a Hilbert space, and $\Gamma \in \mathcal{L}(U;U)$, $\Gamma \geq 0$. We say that the set-valued map $H : U \rightrightarrows U$ is Γ*-strongly monotone at* $\widehat{u}$ *for* $\widehat{w} \in H(\widehat{u})$ if there exists a neighborhood $\mathcal{U} \ni \widehat{u}$ such that for any $u \in \mathcal{U}$ and $w \in H(u)$,

$$\langle w - \widehat{w}, u - \widehat{u}\rangle \geq \|u - \widehat{u}\|_{\Gamma}^2. \tag{3.3}$$

If $\Gamma = 0$, we say that H is monotone at $\widehat{u}$ for $\widehat{w}$.

Assumption 3.3. For any $\widehat{w} = (\widehat{v}, \widehat{\xi}) \in H(\widehat{u})$, the set-valued map ∂G is (Γ_G-strongly) monotone at $\widehat{x}$ for $\widehat{v} - [\nabla K(\widehat{x})]^*\widehat{y}$ in the neighborhood $\mathcal{X}_G$, and the set-valued map ∂F^* is (Γ_{F^*}-strongly) monotone at $\widehat{y}$ for $\widehat{\xi} + K(\widehat{x})$ in the neighborhood $\mathcal{Y}_{F^*}$.

In view of the assumed convexity of G and F^*, Assumption 3.3 is always satisfied with $\Gamma_G = \Gamma_{F^*} = 0$. Also note that the monotonicity of the set-valued map H is closely related to its subregularity [23]: in fact, the former provides an alternative pathway compared to the metric regularity (Aubin property of H^{-1}) employed in [13, 21]. While the discussion of these relationships is beyond the scope of this paper, interested readers are referred to [18, 20] as well as the works discussing strong metric subregularity [1, 10, 15], directional subregularity [17], and partial strong submonotonicity [23].

Combining Assumptions 3.1 and 3.3, throughout the rest of the paper, we assume the neighborhood $\mathcal{U}(\rho_x, \rho_y)$ of $\widehat{u}$ to be nonempty and defined for some $\rho_x, \rho_y > 0$ as

$$\mathcal{U}(\rho_x, \rho_y) := (\mathbb{B}(\widehat{x}, \rho_x) \cap \mathcal{X}_G \cap \mathcal{X}_K) \times (\mathbb{B}_{\mathrm{NL}}(\widehat{y}, \rho_y) \cap \mathcal{Y}_{F^*}). \tag{3.4}$$

3.2 GENERAL ESTIMATES

We verify the conditions of Theorem 2.1 in several steps. First, we ensure that the operator $Z_{i+1}M_{i+1}$ giving rise to the local metric is self-adjoint. Then we show that $Z_{i+2}M_{i+2}$ and the update $Z_{i+1}(M_{i+1} + \Xi_{i+1})$ performed by the algorithm give the same norms (metrics). Here Ξ_{i+1} represents some off-diagonal components from the algorithm, as well as any strong monotonicity available for acceleration. Finally, we estimate $V'_{i+1}(u)$ and $H(u)$ to derive Δ_{i+1}.

We require the following relationships for some $\kappa \in [0, 1)$, $\eta_i > 0$, $\widetilde{\Gamma}_G \in \mathcal{L}(X;X)$, and $\widetilde{\Gamma}_{F^*} \in \mathcal{L}(Y;Y)$:

$$\omega_i := \eta_i/\eta_{i+1}, \qquad \Psi_i\Sigma_i = \eta_i I, \tag{3.5a}$$

$$\Phi_i T_i = \eta_i I, \qquad (1-\kappa)\Psi_{i+1} \geq \eta_i^2 \nabla K(x^i)\Phi_i^{-1}[\nabla K(x^i)]^*, \tag{3.5b}$$

$$\Phi_i = \Phi_i^* \geq 0, \qquad \Psi_{i+1} = \Psi_{i+1}^* \geq 0, \tag{3.5c}$$

$$\Phi_{i+1} = \Phi_i(1 + 2T_i\widetilde{\Gamma}_G), \qquad \Psi_{i+2} = \Psi_{i+1}(1 + 2\Sigma_{i+1}\widetilde{\Gamma}_{F^*}). \tag{3.5d}$$

In Section 4 we will verify these relationships for specific rules for scalar step lengths.

Lemma 3.2. *Fix $i \in \mathbb{N}$ and suppose (3.5) holds. Then $Z_{i+1}M_{i+1}$ is self-adjoint and $Z_{i+1}M_{i+1} \geq \begin{pmatrix} \delta\Phi_i & 0 \\ 0 & (\kappa-\delta)(1-\delta)^{-1}\Psi_{i+1} \end{pmatrix}$ for any $\delta \in [0, \kappa]$.*

Proof. From (2.3) and (3.5), we have $\Phi_i T_i = \eta_i I$ and $\Psi_{i+1}\Sigma_{i+1}\omega_i = \eta_i I$, so that

$$Z_{i+1}M_{i+1} = \begin{pmatrix} \Phi_i & -\eta_i[\nabla K(x^i)]^* \\ -\eta_i \nabla K(x^i) & \Psi_{i+1} \end{pmatrix}. \tag{3.6}$$

Therefore, $Z_{i+1}M_{i+1}$ is self-adjoint. By Cauchy's inequality, also

$$Z_{i+1}M_{i+1} \geq \begin{pmatrix} \delta\Phi_i & 0 \\ 0 & \Psi_{i+1} - \frac{\eta_i^2}{1-\delta}\nabla K(x^i)\Phi_i^{-1}[\nabla K(x^i)]^* \end{pmatrix}. \tag{3.7}$$

Now (3.5) ensures the remaining part of the statement. □

Our next step is to simplify $Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}$ in (CI) while keeping the option to accelerate the method when some of the constituents of H exhibit strong monotonicity.

Lemma 3.3. *Fix $i \in \mathbb{N}$, and suppose (3.5) holds. Then $\frac{1}{2}\|\cdot\|^2_{Z_{i+1}(M_{i+1}+\Xi_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*}))-Z_{i+2}M_{i+2}} = 0$ for*

$$\Xi_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*}) := \begin{pmatrix} 2T_i\widetilde{\Gamma}_G & 2T_i[\nabla K(x^i)]^* \\ -2\Sigma_{i+1}\nabla K(x^{i+1}) & 2\Sigma_{i+1}\widetilde{\Gamma}_{F^*} \end{pmatrix}.$$

Proof. Let $D_{i+2} := Z_{i+1}(M_{i+1} + \Xi_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*})) - Z_{i+2}M_{i+2}$. We can write

$$D_{i+2} = \begin{pmatrix} 0 & [\eta_{i+1}\nabla K(x^{i+1}) + \eta_i \nabla K(x^i)]^* \\ -\eta_{i+1}\nabla K(x^{i+1}) - \eta_i\nabla K(x^i) & 0 \end{pmatrix}$$

using (3.5) and (3.6). This quickly yields the claim. □

Lemma 3.4. *Suppose Assumption 3.1 and (3.5) hold. For a fixed $i \in \mathbb{N}$, let $\overline{x}^{i+1} \in \mathcal{X}_K$ and $\rho_x, \rho_y \geq 0$ be such that $u^i, u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$. Then for any $\zeta, \beta_1 > 0$ and $\alpha_1 \in [0,1]$ we have the estimate*

$$\begin{aligned} \langle V'_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} - \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}\Xi_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*})} \geq \frac{1}{2}\|u^{i+1} - u^i\|^2_{\hat{Q}_{i+1}} \\ + \|u^{i+1} - \widehat{u}\|^2_{Q_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*})}, \end{aligned} \tag{3.8}$$

where

$$Q_{i+1}(\widetilde{\Gamma}_G,\widetilde{\Gamma}_{F^*}) := \begin{pmatrix} -\eta_i(\widetilde{\Gamma}_G + \zeta I) & -\eta_i[\nabla K(x^{i+1})]^* \\ \eta_{i+1}\nabla K(x^{i+1}) & -\eta_{i+1}\left[\widetilde{\Gamma}_{F^*} + \frac{\alpha_1}{\beta_1}L\rho_x P_{\mathrm{NL}}\right] \end{pmatrix}, \quad \text{and}$$

$$\hat{Q}_{i+1} := \begin{pmatrix} -\frac{\eta_i L}{2}\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{\zeta} + (\omega_i+2)[\![2\rho_y, (\omega_i+2)\omega_i\beta_1\rho_x]\!]^{\alpha_1}\right) I & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof. From (2.2) and (3.5), we have

$$
\begin{aligned}
D &:= \langle V'_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} - \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}\Xi_{i+1}(0,0)} \\
&= \eta_i \langle [\nabla K(x^i) - \nabla K(x^{i+1})](x^{i+1} - \widehat{x}), y^{i+1}\rangle \\
&\quad + \eta_{i+1}\langle K(x^{i+1}) - K(\bar{x}^{i+1}) - \nabla K(x^i)(x^{i+1} - \bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle \\
&\quad + \langle (\eta_{i+1}\nabla K(x^{i+1}) - \eta_i \nabla K(x^i))(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle.
\end{aligned}
\tag{3.9}
$$

Rearranging the terms, we obtain

$$
\begin{aligned}
D = \eta_i &\langle [\nabla K(x^i) - \nabla K(x^{i+1})](x^{i+1} - \widehat{x}), \widehat{y}\rangle \\
&+ \eta_{i+1}\langle K(x^{i+1}) - K(\bar{x}^{i+1}) - \nabla K(x^{i+1})(x^{i+1} - \bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle \\
&+ \eta_{i+1}\langle (\nabla K(x^{i+1}) - \nabla K(x^i))(x^{i+1} - \bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle \\
&+ (\eta_{i+1} - \eta_i)\langle \nabla K(x^{i+1})(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle.
\end{aligned}
$$

Using (3.5) and the Lipschitz property of Assumption 3.1, we further estimate

$$
\begin{aligned}
D \geq \eta_i &\langle [\nabla K(x^i) - \nabla K(x^{i+1})](x^{i+1} - \widehat{x}), \widehat{y}\rangle - \eta_{i+1}(L/2)\|x^{i+1} - \bar{x}^{i+1}\|^2 \|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \\
&- \eta_{i+1} L \|x^{i+1} - x^i\| \|x^{i+1} - \bar{x}^{i+1}\| \|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \\
&+ (\eta_{i+1} - \eta_i)\langle \nabla K(x^{i+1})(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle.
\end{aligned}
$$

Since $\bar{x}^{i+1} - x^{i+1} = \omega_i(x^{i+1} - x^i)$, using (3.5) we obtain

$$
\begin{aligned}
D \geq \eta_i &\langle [\nabla K(x^i) - \nabla K(x^{i+1})](x^{i+1} - \widehat{x}), \widehat{y}\rangle \\
&- \eta_i L(1 + \omega_i/2)\|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \|x^{i+1} - x^i\|^2 \\
&+ (\eta_{i+1} - \eta_i)\langle \nabla K(x^{i+1})(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle.
\end{aligned}
\tag{3.10}
$$

To later allow balancing between further assumptions on the primal and dual, we pick any $\alpha_1 \in [0, 1]$, and multiply the middle term by $1 = \alpha_1 + (1 - \alpha_1)$. We then apply Cauchy's inequality on the part weighted by α_1, as well as Assumption 3.1 and Cauchy's inequality on the first term of (3.10), to obtain for any $\zeta, \beta_1 > 0$ the estimate

$$
\begin{aligned}
D \geq &-\frac{\eta_i L^2 \|P_{\mathrm{NL}}\widehat{y}\|^2}{4\zeta}\|x^{i+1} - x^i\|^2 + (\eta_{i+1} - \eta_i)\langle \nabla K(x^{i+1})(x^{i+1} - \widehat{x}), y^{i+1} - \widehat{y}\rangle \\
&- \eta_i \zeta \|x^{i+1} - \widehat{x}\|^2 - \eta_i(1 - \alpha_1)L(1 + \omega_i/2)\|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \|x^{i+1} - x^i\|^2 \\
&- \eta_i \alpha_1 L \|x^{i+1} - x^i\| \left(\frac{1}{2\beta_1\omega_i}\|y^{i+1} - \widehat{y}\|^2_{P_{\mathrm{NL}}} + \frac{(\omega_i + 2)^2 \omega_i \beta_1}{8}\|x^{i+1} - x^i\|^2 \right).
\end{aligned}
$$

Using $\|x^{i+1} - x^i\| \leq 2\rho_x$ and $\|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}} \leq \rho_y$, we finally get $D \geq (1/2)\|u^{i+1} - u^i\|^2_{\hat{Q}_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{Q_{i+1}(0,0)}$, in which the right-hand side differs from that of (3.8) by having $Q_{i+1}(0, 0)$ in place of $Q_{i+1}(\widetilde{\Gamma}_G, \widetilde{\Gamma}_{F^*})$. Recalling how D is defined in (3.9), we may add back the difference to obtain the claim. ◻

We now proceed to the final steps necessary for the Δ_{i+1} estimate.

Lemma 3.5. *Suppose $\widehat{w} = (\widehat{v}, \widehat{\xi}) \in H(\widehat{u})$, and that Assumptions 3.1 and 3.3 hold. Let $\rho_x, \rho_y > 0$, as well as $\beta_2 > 0$ and $\alpha_2 \in [0,1]$. Define*

$$\Gamma_{H,i+1}(u) := \begin{pmatrix} \eta_i\left[\Gamma_G + \Theta - \frac{L}{16\omega_i}[\![8\rho_y, \beta_2\rho_x]\!]^{\alpha_2} I\right] & \eta_i[\nabla K(x)]^* \\ -\eta_{i+1}\nabla K(x) & \eta_{i+1}\left[\Gamma_{F^*} - \frac{\alpha_2}{\beta_2} L\rho_x P_{\mathrm{NL}}\right] \end{pmatrix}.$$

Then for all $u \in \mathcal{U}(\rho_x, \rho_y)$ holds

$$\langle H(u) - \widehat{w}, u - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} \geq \|u - \widehat{u}\|^2_{\Gamma_{H,i+1}(u)}. \tag{3.11}$$

Proof. Since $\widehat{w} \in H(\widehat{u})$, we have $\partial G(\widehat{x}) \ni z_G := \widehat{v} - [\nabla K(\widehat{x})]^*\widehat{y}$, and $\partial F^*(\widehat{y}) \ni z_{F^*} := \widehat{\xi} + K(\widehat{x})$. Using (3.5), we therefore expand

$$\begin{aligned}\langle H(u) - \widehat{w}, u - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} &= \eta_i\langle \partial G(x) - z_G, x - \widehat{x}\rangle + \eta_{i+1}\langle \partial F^*(y) - z_{F^*}, y - \widehat{y}\rangle \\ &\quad + \eta_i\langle [\nabla K(x)]^* y - [\nabla K(\widehat{x})]^*\widehat{y}, x - \widehat{x}\rangle + \eta_{i+1}\langle K(\widehat{x}) - K(x), y - \widehat{y}\rangle.\end{aligned}$$

Using the local (strong) monotonicity of G and F^* and rearranging terms, we obtain

$$\begin{aligned}\langle H(u) - \widehat{w}, u - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} &\geq \eta_i\|x - \widehat{x}\|^2_{\Gamma_G} + \eta_{i+1}\|y - \widehat{y}\|^2_{\Gamma_{F^*}} \\ &\quad + \eta_i\langle(\nabla K(x) - \nabla K(\widehat{x}))(x - \widehat{x}), \widehat{y}\rangle + (\eta_i - \eta_{i+1})\langle \nabla K(x)(x - \widehat{x}), y - \widehat{y}\rangle \\ &\quad + \eta_{i+1}\langle K(\widehat{x}) - K(x) + \nabla K(x)(x - \widehat{x}), y - \widehat{y}\rangle.\end{aligned} \tag{3.12}$$

Using both the Lipschitz property and hypomonotonicity of Assumption 3.1 we obtain

$$\begin{aligned}\langle H(u) - \widehat{w}, u - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} &\geq \eta_i\|x - \widehat{x}\|^2_{\Gamma_G + \Theta} + \eta_{i+1}\|y - \widehat{y}\|^2_{\Gamma_{F^*}} \\ &\quad - \eta_{i+1}(L/2)\|x - \widehat{x}\|^2\|y - \widehat{y}\|_{P_{\mathrm{NL}}} + (\eta_i - \eta_{i+1})\langle \nabla K(x)(x - \widehat{x}), y - \widehat{y}\rangle.\end{aligned}$$

Similarly to the proof of Lemma 3.4, to allow balancing between primal and dual assumptions in the future, we multiply the middle term by $1 = \alpha_2 + (1 - \alpha_2)$. We then apply Cauchy's inequality to the part multiplied by α_2 to obtain for our choice of $\beta_2 > 0$ the estimate

$$\begin{aligned}\langle H(u) - \widehat{w}, u - \widehat{u}\rangle_{Z_{i+1}W_{i+1}} &\geq \eta_i\|x - \widehat{x}\|^2_{\Gamma_G + \Theta} + \eta_{i+1}\|y - \widehat{y}\|^2_{\Gamma_{F^*}} \\ &\quad - \eta_{i+1}(1 - \alpha_2)(L/2)\|x - \widehat{x}\|^2\|y - \widehat{y}\|_{P_{\mathrm{NL}}} \\ &\quad - \eta_{i+1}\alpha_2 L\|x - \widehat{x}\|\left(\frac{1}{\beta_2}\|y - \widehat{y}\|^2_{P_{\mathrm{NL}}} + \frac{\beta_2}{16}\|x - \widehat{x}\|^2\right) \\ &\quad + (\eta_i - \eta_{i+1})\langle \nabla K(x)(x - \widehat{x}), y - \widehat{y}\rangle.\end{aligned} \tag{3.13}$$

Rearranging terms gives (3.11). □

We now have all the necessary tools to formulate the main estimate. Combining the results of the previous lemmas, we arrive at the following conclusion:

Theorem 3.6. *Fix $i \in \mathbb{N}$, and suppose (3.5) and Assumptions 3.1 and 3.3 hold. Also suppose $\bar{x}^{i+1} \in \mathcal{X}_K$, and that $\rho_x, \rho_y \geq 0$ are such that $u^i, u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$. Then (CI) is satisfied (for this i) if*

$$\frac{1}{2}\|u^{i+1} - u^i\|^2_{S_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{\hat{S}_{i+1}} \geq -\Delta_{i+1},$$

where for some $0 \leq \delta \leq \kappa < 1$; $\zeta, \beta_1, \beta_2 > 0$; and $\alpha_1, \alpha_2 \in [0, 1]$ we define

$$S_{i+1} := \begin{pmatrix} \delta\Phi_i - \frac{\eta_i L}{2}\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{\zeta} + (\omega_i+2)[\![2\rho_y, (\omega_i+2)\omega_i\beta_1\rho_x]\!]^{\alpha_1}\right) I & 0 \\ 0 & \Psi_{i+1} - \frac{\eta_i^2}{1-\kappa}\nabla K(x^i)\Phi_i^{-1}[\nabla K(x^i)]^* \end{pmatrix},$$

$$\hat{S}_{i+1} := \begin{pmatrix} \eta_i\left[\Gamma_G - \widetilde{\Gamma}_G + \Theta - \left(\zeta + \frac{L}{16\omega_i}[\![8\rho_y, \beta_2\rho_x]\!]^{\alpha_2}\right) I\right] & 0 \\ 0 & \eta_{i+1}[\Gamma_{F^*} - \widetilde{\Gamma}_{F^*} - \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right) L\rho_x P_{\mathrm{NL}}] \end{pmatrix}.$$

We may in particular take $\Delta_{i+1} = 0$ in (CI) provided

$$\Phi_i \geq \frac{\eta_i L}{2\delta}\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{\zeta} + (\omega_i + 2)[\![2\rho_y, (\omega_i + 2)\omega_i\beta_1\rho_x]\!]^{\alpha_1}\right) I, \tag{3.14a}$$

$$\Psi_{i+1} \geq \frac{\eta_i^2}{1-\kappa}\nabla K(x^i)\Phi_i^{-1}[\nabla K(x^i)]^*, \tag{3.14b}$$

$$\Gamma_G + \Theta \geq \widetilde{\Gamma}_G + \left(\zeta + \frac{L}{16\omega_i}[\![8\rho_y, \beta_2\rho_x]\!]^{\alpha_2}\right) I, \quad \textit{and} \tag{3.14c}$$

$$\Gamma_{F^*} \geq \widetilde{\Gamma}_{F^*} + \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right) L\rho_x P_{\mathrm{NL}}. \tag{3.14d}$$

Proof. Applying Lemma 3.3 to the left-hand side of (CI), we obtain

$$\begin{aligned}\Delta &:= \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}} + \langle \widetilde{H}_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} \\ &= \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \langle V'_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} - \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}\Xi_{i+1}(\widetilde{\Gamma}_G, \widetilde{\Gamma}_{F^*})} \\ &\quad + \langle H(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}}.\end{aligned}$$

Applying Lemma 3.4 and Lemma 3.5, further

$$\Delta \geq \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1} + \hat{Q}_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{Q_{i+1}(\widetilde{\Gamma}_G, \widetilde{\Gamma}_{F^*}) + \Gamma_{H,i+1}(u^{i+1})}.$$

After applying Lemma 3.2 and rearranging terms, we obtain the claim. □

Discussion While (3.14a) and (3.14b) appear to bound Φ_i and Ψ_{i+1}, they, in fact, bound the step lengths. Recall from (3.5) that $\eta_i I = \Phi_i T_i = \Psi_i \Sigma_i$. Therefore, Φ_i and Ψ_{i+1} can be made to vanish, as we will do in Section 4 for scalar step lengths.

The parameter ζ was introduced to estimate (3.10). In Section 4.5, we eliminate ζ when the gradient of $x \mapsto \langle K(x), \widehat{y}\rangle$ is hypomonotone. Otherwise, the best bound for T or Φ in (3.14a) is obtained by choosing the maximal ζ satisfying (3.14c).

If K is linear, as in [7], (3.14a) reduces to $\Phi_i \geq 0$ via $\|P_{\mathrm{NL}}\widehat{y}\| = \rho_y = 0$. Then we can set $\kappa = 0$, so that (3.14b) turns into an operator analogue of the step length bound $\tau_i\sigma_i\|K\|^2 < 1$ of [7]; see also [22]. We can also set $\zeta = 0$ for linear K.

The inequalities (3.14) also imply $\Gamma_G + \Theta > 0$. This was not required in the main result of [21], but the verification of metric regularity for specific problems in [21, Proposition 4.2] would introduce a similar condition. In general, we do not require $\Theta \geq 0$ as long as any negativity is compensated for by the strong convexity of G.

Another difference from [21] is (3.14d): Γ_{F^*} is allowed to be zero, so we do not require strong convexity from F^*; see also [12]. Indeed, α_1 and α_2 allow balancing between small ρ_y but no strong convexity of F^*, similar to [13], and less restrictions on ρ_y, but strong convexity of F^* and a small ρ_x. Thus, the above two alternatives, which we analyze in further detail in Section 4, resemble those in [12, §2.1 and §2.2].

3.3 LOCAL STEP LENGTH BOUNDS

One final technical result needed for convergence estimates is to ensure that $u^{i+1} \in \mathcal{U}(\rho_x, \rho_y)$ once $u^i \in \mathcal{U}(\rho_x, \rho_y)$, as required by Lemma 3.4, Lemma 3.5, and, consequently, Theorem 3.6. The following lemma provides the basis from which we further work in Section 4.2.

Lemma 3.7. *Fix $i \in \mathbb{N}$. Suppose Assumption 3.1 holds and u^{i+1} solves (PP). For simplicity, assume $\omega_i \leq 1$. For some $r_{x,i}, r_{y,i} > 0$, and $\delta_x, \delta_y \in (0,1)$, let $\mathbb{B}(\widehat{x}, r_{x,i}) \subset \mathcal{X}_K$, $x^i \in \mathbb{B}(\widehat{x}, \delta_x r_{x,i})$, and $y^i \in \mathbb{B}(\widehat{y}, \delta_y r_{y,i})$. Then $x^{i+1}, \bar{x}^{i+1} \in \mathbb{B}(\widehat{x}, r_{x,i})$ and $y^{i+1} \in \mathbb{B}(\widehat{y}, r_{y,i})$ provided*

$$\|T_i\| \leq \frac{(1-\delta_x)r_{x,i}}{2\|\nabla K(x^i)\|r_{y,i} + 2L\|P_{\mathrm{NL}}\widehat{y}\|r_{x,i}}, \quad \textit{and} \quad \|\Sigma_{i+1}\| \leq \frac{2(1-\delta_y)r_{y,i}}{(Lr_{x,i} + 2\|\nabla K(\widehat{x})\|)r_{x,i}}. \tag{3.15}$$

Proof. We want to show that the step length conditions (3.15) are sufficient for

$$\|x^{i+1} - \widehat{x}\| \leq r_{x,i}, \quad \|\bar{x}^{i+1} - \widehat{x}\| \leq r_{x,i}, \quad \text{and} \quad \|y^{i+1} - \widehat{y}\| \leq r_{y,i}.$$

We do this by applying the testing argument on the primal and dual variables separately. Multiplying (PP) by $Z_{i+1}^*(u^{i+1} - \widehat{u})$ with $\Phi_i = I$ and $\Psi_{i+1} = 0$, we get

$$0 \in \langle \partial G(x^{i+1}) + [\nabla K(x^i)]^* y^i, x^{i+1} - \widehat{x}\rangle_{T_i} + \langle x^{i+1} - x^i, x^{i+1} - \widehat{x}\rangle.$$

Using the standard three-point formula or Pythagoras' identity

$$\langle x^{i+1} - x^i, x^{i+1} - \widehat{x}\rangle = \frac{1}{2}\|x^{i+1} - x^i\|^2 - \frac{1}{2}\|x^i - \widehat{x}\|^2 + \frac{1}{2}\|x^{i+1} - \widehat{x}\|^2, \tag{3.16}$$

we obtain

$$\|x^i - \widehat{x}\|^2 \in 2\langle \partial G(x^{i+1}) + [\nabla K(x^i)]^* y^i, x^{i+1} - \widehat{x}\rangle_{T_i} + \|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2.$$

Using $0 \in \partial G(\widehat{x}) + [\nabla K(\widehat{x})]^*\widehat{y}$ and the monotonicity of ∂G, we then arrive at

$$\|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2 + 2\langle [\nabla K(x^i)]^* y^i - [\nabla K(\widehat{x})]^*\widehat{y}, x^{i+1} - \widehat{x}\rangle_{T_i} \leq \|x^i - \widehat{x}\|^2.$$

With $C_x := \|[\nabla K(x^i)]^* y^i - [\nabla K(\widehat{x})]^* \widehat{y}\|_{T_i^2}$ then

$$\|x^{i+1} - x^i\|^2 + \|x^{i+1} - \widehat{x}\|^2 \le 2C_x\|x^{i+1} - \widehat{x}\| + \|x^i - \widehat{x}\|^2, \tag{3.17}$$

or, after rearranging the terms and using $\|x^{i+1} - \widehat{x}\| \le \|x^{i+1} - x^i\| + \|x^i - \widehat{x}\|$,

$$(\|x^{i+1} - x^i\| - C_x)^2 + \|x^{i+1} - \widehat{x}\|^2 \le (\|x^i - \widehat{x}\| + C_x)^2,$$

which leads to

$$\|x^{i+1} - \widehat{x}\| \le \|x^i - \widehat{x}\| + C_x. \tag{3.18}$$

Hence, if $C_x \le (1 - \delta_x) r_{x,i}$, we get the first required estimate $\|x^{i+1} - \widehat{x}\| \le r_{x,i}$.

To estimate the dual variable, we multiply (PP) by $Z_{i+1}^*(u^{i+1} - \widehat{u})$ with $\Phi_i = 0$, $\Psi_{i+1} = I$. This gives

$$0 \in \langle \partial F^*(y^{i+1}) - K(\bar{x}^{i+1}), y^{i+1} - \widehat{y}\rangle_{\Sigma_{i+1}} + \langle y^{i+1} - y^i, y^{i+1} - \widehat{y}\rangle.$$

Using $0 \in \partial F^*(\widehat{y}) - K(\widehat{x})$ and following the steps leading to (3.18), we deduce

$$\|y^{i+1} - \widehat{y}\| \le \|y^i - \widehat{y}\| + C_y \tag{3.19}$$

with $C_y := \|K(\widehat{x}) - K(\bar{x}^{i+1})\|_{\Sigma_{i+1}^2}$. Consequently, if $C_y \le (1 - \delta_y) r_{y,i}$, then $\|y^{i+1} - \widehat{y}\| \le r_{y,i}$.

We now proceed to deriving bounds on C_x and C_y with the goal of bounding (3.18) and (3.19) from above. Using Assumption 3.1, and arguing as in (3.2), we estimate

$$C_x \le \|T_i\|(\|\nabla K(x^i)\|\|y^i - \widehat{y}\| + L\|P_{\mathrm{NL}}\widehat{y}\|\|x^i - \widehat{x}\|) =: R_x, \quad \text{and} \tag{3.20}$$

$$C_y \le \|\Sigma_{i+1}\|(L\|\bar{x}^{i+1} - \widehat{x}\|/2 + \|\nabla K(\widehat{x})\|)\|\bar{x}^{i+1} - \widehat{x}\| =: R_y \quad (\text{if } \bar{x}^{i+1} \in \mathcal{X}_K). \tag{3.21}$$

We need to verify that $\bar{x}^{i+1} \in \mathcal{X}_K$, used for the bound on C_y. By definition,

$$\begin{aligned}
\|\bar{x}^{i+1} - \widehat{x}\|^2 &= \|x^{i+1} - \widehat{x} + \omega_i(x^{i+1} - x^i)\|^2 \\
&= \|x^{i+1} - \widehat{x}\|^2 + \omega_i^2\|x^{i+1} - x^i\|^2 + 2\omega_i\langle x^{i+1} - \widehat{x}, x^{i+1} - x^i\rangle \\
&= (1 + \omega_i)\|x^{i+1} - \widehat{x}\|^2 + \omega_i(1 + \omega_i)\|x^{i+1} - x^i\|^2 - \omega_i\|x^i - \widehat{x}\|^2 \\
&\le (1 + \omega_i)(\|x^{i+1} - \widehat{x}\|^2 + \|x^{i+1} - x^i\|^2) - \omega_i\|x^i - \widehat{x}\|^2.
\end{aligned}$$

Applying (3.17) and (3.18), we obtain

$$\|\bar{x}^{i+1} - \widehat{x}\|^2 \le 4C_x\|x^{i+1} - \widehat{x}\| + \|x^i - \widehat{x}\|^2 \le 4C_x(\delta_x r_{x,i} + C_x) + \delta_x^2 r_{x,i}^2.$$

Hence, $\|\bar{x}^{i+1} - \widehat{x}\| \le r_{x,i}$ if

$$4C_x(\delta_x r_{x,i} + C_x) + \delta_x^2 r_{x,i}^2 \le r_{x,i}^2 \iff C_x \le (1 - \delta_x) r_{x,i}/2.$$

Consequently, if $R_x \le (1 - \delta_x) r_{x,i}/2$, then $\|\bar{x}^{i+1} - \widehat{x}\| \le r_{x,i}$. Coincidentally, from (3.18) we get $\|x^{i+1} - \widehat{x}\| \le r_{x,i}$. Then we impose $R_y \le (1 - \delta_y) r_{y,i}$ so that (3.19) yields $\|y^{i+1} - \widehat{y}\| \le r_{y,i}$. After substituting the expression for R_x and R_y in (3.20) and (3.21), these imposed bounds expand into

$$\|T_i\| \le \frac{(1 - \delta_x) r_{x,i}}{2\|\nabla K(x^i)\|\delta_y r_{y,i} + 2L\|P_{\mathrm{NL}}\widehat{y}\|\delta_x r_{x,i}}, \quad \text{and} \quad \|\Sigma_{i+1}\| \le \frac{(1 - \delta_y) r_{y,i}}{\left(L r_{x,i}/2 + \|\nabla K(\widehat{x})\|\right) r_{x,i}}.$$

Since $\delta_x, \delta_y < 1$, the bounds from the statement of the lemma will also suffice. □

Remark 3.8. *If we can take $\mathcal{X}_K = X$ in Assumption 3.1, $\delta_x r_{x,i}$ will stand for $\|x^i - \widehat{x}\|$, and the upper bound on $\|T_i\|$ in Lemma 3.7 will only be needed to ensure $y^{i+1} \in \mathbb{B}(\widehat{y}, r_{y,i})$. However, if $r_{x,i}$ escapes to infinity, the constraint on $\|\Sigma_{i+1}\|$ in (3.15) goes to zero. Therefore, other approaches are needed to ensure $y^{i+1} \in \mathbb{B}(\widehat{y}, r_{y,i})$.*

In particular, $r_{x,i} \to \infty$ is not a problem if $\operatorname{dom} F^*$ is bounded and $\operatorname{dom} F^* \subseteq \mathbb{B}(\widehat{y}, r_{y,i})$. Indeed the operator $(I + \sigma\partial F^*)^{-1}$ in Algorithm 1.1 will always ensure $y^{i+1} \in \operatorname{dom} F^*$. Hence, if $\mathcal{X}_K = X$ and $\operatorname{dom} F^* \subseteq \mathbb{B}(\widehat{y}, r_{y,i})$, $u^{i+1} \in \mathcal{X}_K \times \mathcal{Y}$ always, unconditionally. Moreover, if $\alpha_1 = \alpha_2 = 0$ in (3.14), the step length bounds will only depend on ρ_y. Consequently, it will suffice to verify $\operatorname{dom} F^* \subseteq \mathbb{B}_{\mathrm{NL}}(\widehat{y}, \rho_y)$.

4 REFINEMENT TO SCALAR STEP LENGTHS

To derive convergence rates, we now simplify Theorem 3.6 for scalar step lengths. Specifically, we assume for some scalars $\gamma_G, \gamma_{F^*}, \tau_i, \phi_i, \sigma_i, \psi_i \geq 0$, and $\theta \in \mathbb{R}$, the structure

$$
\begin{cases} T_i := \tau_i I, & \Phi_i := \phi_i I, \quad \Gamma_G := \gamma_G I, \\ \Sigma_i := \sigma_i I, & \Psi_i := \psi_i I, \quad \Gamma_{F^*} := \gamma_{F^*} I, \quad \text{and} \quad \Theta := \theta I. \end{cases} \tag{4.1}
$$

This reduces (PP) to Algorithm 1.1, which for convex, proper, lower semicontinuous G and F^* is always solvable for the iterates $\{u^i := (x^i, y^i)\}_{i\in\mathbb{N}}$.

For the sake of brevity and simplicity, we divide our analysis into the two cases $\alpha_1 = \alpha_2 = 0$ and $\alpha_1 = \alpha_2 = 1$, in the respective Sections 4.3 and 4.4. We explain the implications of these choices in Section 4.1. In both cases, we show weak and strong convergence for constant step lengths, and provide step length rules that ensure $O(1/N^2)$ under primal strong monotonicity, and linear convergence under primal–dual strong monotonicity. Finally, in Section 4.5, we consider the particular case of $\langle K(\cdot), \widehat{y}\rangle$ having a hypomonotone gradient.

4.1 GENERAL DERIVATIONS AND ASSUMPTIONS

Under the setup (4.1), the rules (3.5) and (3.14) demand for some $\alpha_1, \alpha_2 \in [0, 1]$; $\beta_1, \beta_2, \zeta > 0$; $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*} \geq 0$ (non-negativity introduced here); and $0 \leq \delta \leq \kappa < 1$ that

$$
\omega_i = \eta_i/\eta_{i+1}, \qquad \eta_i = \psi_i\sigma_i = \phi_i\tau_i, \tag{4.2a}
$$

$$
\phi_{i+1} = \phi_i(1 + 2\tau_i\widetilde{\gamma}_G), \qquad \psi_{i+2} = \psi_{i+1}(1 + 2\sigma_{i+1}\widetilde{\gamma}_{F^*}), \tag{4.2b}
$$

$$
\psi_{i+1} \geq \frac{\eta_i^2\phi_i^{-1}}{1-\kappa}\|\nabla K(x^i)\|^2, \qquad \gamma_{F^*} \geq \widetilde{\gamma}_{F^*} + \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right)L\rho_x, \tag{4.2c}
$$

$$
\gamma_G + \theta \geq \widetilde{\gamma}_G + \left(\zeta + \frac{L}{16\omega_i}[\![8\rho_y, \beta_2\rho_x]\!]^{\alpha_2}\right), \qquad \text{and} \tag{4.2d}
$$

$$
\phi_i \geq \frac{\eta_i L}{2\delta}\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{\zeta} + (\omega_i + 2)[\![2\rho_y, (\omega_i + 2)\omega_i\beta_1\rho_x]\!]^{\alpha_1}\right). \tag{4.2e}
$$

If $\alpha_1 = \alpha_2 = 0$, (4.2) will not depend on ρ_x. Indeed, substituting $\eta_i = \phi_i\tau_i$ and $\eta_i^2 = \phi_i\tau_i\psi_i\sigma_i$

in (4.2e) and the first part of (4.2c), we get the upper bounds

$$\tau_i \le \frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2\zeta} + (\omega_i+2)\,\rho_y\right)}, \quad \text{and} \quad \sigma_i\tau_i \le \frac{1-\kappa}{R_K^2}, \tag{4.3}$$

where $R_K = \sup_{\mathcal{X}} \|\nabla K(x)\|$. In the latter bound, we also used $\psi_{i+1} \ge \psi_i$, which follows from assumed (4.2b) and $\gamma_{F^*} \ge 0$. We see from (4.3) that a level of dual locality is required: we need the bound ρ_y for τ_i to be finite. We consider this case in detail in Section 4.3

If $\alpha_1 = \alpha_2 = 1$, (4.2) does not depend on ρ_y. To satisfy the second inequality in (4.2c) and (4.2d), we select $\beta_1 = L\rho_x\beta_2/((\gamma_{F^*} - \widetilde{\gamma}_{F^*})\beta_2 - L\rho_x)$ and $\zeta = \gamma_G + \theta - \widetilde{\gamma}_G - L\beta_2\rho_x/16\omega_i$. If we define $\hat{\gamma}_G := (\gamma_G + \theta - \widetilde{\gamma}_G)/L$ and $\hat{\gamma}_{F^*} := (\gamma_{F^*} - \widetilde{\gamma}_{F^*})/L$, the required nonnegativity of β_1 and ζ will hold if and only if $\beta_2 \in (\rho_x/\hat{\gamma}_{F^*}, 16\omega_i\hat{\gamma}_G/\rho_x)$. Consequently, for such β_1, β_2, and ζ, we require the following condition, which will be sufficient for (4.2c) and (4.2d) to hold:

$$\rho_x < 4\sqrt{\omega_i\hat{\gamma}_G\hat{\gamma}_{F^*}}. \tag{4.4}$$

For such ρ_x, substituting $\eta_i = \phi_i\tau_i$, β_1, and ζ to (4.2e), we obtain

$$\tau_i \le \frac{2\delta}{L\omega_i\left(\frac{16\|P_{\mathrm{NL}}\widehat{y}\|^2}{16\omega_i\hat{\gamma}_G - \rho_x\beta_2} + \frac{(\omega_i+2)^2\rho_x^2\beta_2}{\hat{\gamma}_{F^*}\beta_2 - \rho_x}\right)}.$$

Minimizing the denominator in β_2 over $[\rho_x/\hat{\gamma}_{F^*}, 16\omega_i\hat{\gamma}_G/\rho_x]$, we arrive at

$$\begin{aligned} &\tau_i \le \frac{\delta(16\omega_i\hat{\gamma}_G\hat{\gamma}_{F^*} - \rho_x^2)}{d_i} \quad \text{and} \quad \sigma_i\tau_i \le \frac{1-\kappa}{R_K^2}, \quad \text{where} \\ &d_i := 4L\omega_i(2\|P_{\mathrm{NL}}\widehat{y}\|^2\hat{\gamma}_{F^*} + \|P_{\mathrm{NL}}\widehat{y}\|(\omega_i+2)\rho_x^2 + 2\omega_i(\omega_i+2)^2\hat{\gamma}_G\rho_x^2). \end{aligned} \tag{4.5}$$

Now, to get useful step lengths, we need a form of primal locality: ρ_x has to be sufficiently small. We also need $\gamma_{F^*} > 0$, i.e., F^* to be strongly convex. We study this case in detail in Section 4.4.

While the bounds above will further be refined in the coming lemmas and theorems, we collect the refinements of all the more structural assumptions of Section 3 in:

Assumption 4.1. Suppose $G : X \to \overline{\mathbb{R}}$ and $F^* \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous; $K \in C^1(X;Y)$; and the following hold for some $\rho_x, \rho_y \ge 0$ and the iterates $\{u^i\}_{i\in\mathbb{N}}$ of Algorithm 1.1:

(i) (*Locally Lipschitz* ∇K) There exists $L \ge 0$ with $\|\nabla K(x) - \nabla K(x')\| \le L\|x - x'\|$ for any $x, x' \in \mathcal{X}_K$;

(ii) (*Locally bounded* ∇K) There exists $R_K > 0$ with $\sup_{x\in\mathcal{X}_K} \|\nabla K(x)\| \le R_K$;

(iii) (*Monotone* ∂G *and* ∂F^*) For any $\widehat{w} = (\widehat{v}, \widehat{\xi}) \in H(\widehat{u})$, the map ∂G is $\gamma_G I$-strongly monotone at $\widehat{x}$ for $\widehat{v} - [\nabla K(\widehat{x})]^*\widehat{y}$ in $\mathcal{X}_G$ with $\gamma_G \ge 0$; and the map ∂F^* is $\gamma_{F^*}I$-strongly monotone at $\widehat{y}$ for $\widehat{\xi} + K(\widehat{x})$ in $\mathcal{Y}_{F^*}$ with $\gamma_{F^*} \ge 0$;

(iv) (*Point-hypomonotone saddle term gradient*) There exists $\theta \in \mathbb{R}$ such that $\langle(\nabla K(x) - \nabla K(\widehat{x}))(x - \widehat{x}), \widehat{y}\rangle \ge \theta\|x - \widehat{x}\|^2$ for any $u \in \mathcal{U}(\rho_x, \rho_y)$ defined by (3.4);

(v) (*Neighborhood-compatible iterations*) $\{u^i\}_{i\in\mathbb{N}} \in \mathcal{U}(\rho_x, \rho_y)$ with $\{\bar{x}^{i+1}\}_{i\in\mathbb{N}} \in \mathcal{X}_K$.

We will not further refine the above assumptions other than providing sufficient conditions for Assumption 4.1(v) in the following subsection.

4.2 NEIGHBORHOOD-COMPATIBLE ITERATIONS

The purpose of this subsection is to provide explicit formulas to ensure Assumption 4.1(v) holds.

Lemma 4.1. *Let $\delta_x, \delta_y \in (0,1)$, as well as $0 \le \delta \le \kappa < 1$ and $\rho_x, \rho_y > 0$ be given, and assume (4.2) holds with $\psi_1 = 1$, $1/\sqrt{1+2\tau_i\widetilde{\gamma}_G} \le \omega_i \le 1$ for all $i \in \mathbb{N}$. Also assume $\sup_{x\in\mathcal{X}_K} \|\nabla K(x)\| \le R_K$. Define*

$$\mu := \sigma_0/\tau_0, \quad r_{\min} := \|x_0 - \widehat{x}\|/\delta_x, \quad \textit{and} \quad r_{\max} := \delta_x^{-1}\sqrt{2\delta^{-1}(\|x_0-\widehat{x}\|^2 + \mu^{-1}\|y_0-\widehat{y}\|^2)}.$$

Assume that $\mathbb{B}(\widehat{x}, r_{\max}) \times \mathbb{B}(\widehat{y}, r_y) \subseteq \mathcal{U}(\rho_x, \rho_y)$ for some $r_y \ge r_{\max}\sqrt{\mu/(\kappa-\delta)}/(2\delta_y)$ and the step length τ_0 to satisfy

$$\tau_0 \le \min\left\{\frac{(1-\delta_x)r_{\min}}{2R_K r_y + 2L\|P_{\mathrm{NL}}\widehat{y}\|r_{\min}}, \frac{2(1-\delta_y)\omega_0 r_y}{(Lr_{\max}+2R_K)r_{\max}\mu}\right\}. \tag{4.6}$$

Then Assumption 4.1(v) holds.

Proof of Lemma 4.1. The proof will be carried out by induction. We will show that $u^i \in \mathbb{B}(\widehat{x}, \delta_x r_{x,i}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$, and that

$$\frac{\tau_i}{r_{x,i}} \le \frac{1-\delta_x}{2R_K r_y + 2L\|P_{\mathrm{NL}}\widehat{y}\|r_{x,i}}, \quad \text{and} \quad \sigma_{i+1}r_{x,i} \le \frac{2(1-\delta_y)r_y}{Lr_{x,i}+2R_K}, \tag{4.7}$$

where $r_{x,i} := \|u_0 - \widehat{u}\|_{Z_1M_1}/(\sqrt{\delta\phi_i}\delta_x)$. For this purpose, we introduce the sets

$$\mathcal{U}_i := \left\{(x,y) \in X \times Y \;\middle|\; \|x-\widehat{x}\|^2 + \tfrac{\psi_{i+1}}{\phi_i}\tfrac{\kappa-\delta}{(1-\delta)\delta}\|y-\widehat{y}\|^2 \le \delta_x^2 r_{x,i}^2\right\}$$

and show that $\mathcal{U}_i \subseteq \mathbb{B}(\widehat{x}, \delta_x r_{x,i}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$. It immediately follows that $\mathcal{U}_i \subseteq \mathbb{B}(\widehat{x}, \delta_x r_{x,i}) \times Y$. To show $\mathcal{U}_i \subseteq X \times \mathbb{B}(\widehat{y}, \delta_y r_y)$, it is enough if we demonstrate

$$\delta_y^2 r_y^2 \ge \frac{\delta_x^2 r_{x,i}^2\phi_i}{\psi_{i+1}}\frac{(1-\delta)\delta}{\kappa-\delta} = \frac{\delta_x^2 r_{x,0}^2\phi_0}{\psi_{i+1}}\frac{(1-\delta)\delta}{\kappa-\delta}. \tag{4.8}$$

Due to (4.2), $\widetilde{\gamma}_{F^*} \ge 0$ and therefore $\psi_{i+1} \ge \psi_i \ge \psi_1 \ge 1$ as well as $\phi_0 = \mu$. If we expand $r_{x,0}$ and apply Lemma 3.2, we obtain

$$r_{x,0}^2 = \|u_0 - \widehat{u}\|_{Z_1M_1}^2/(\delta\phi_0\delta_x^2) \ge \|x_0-\widehat{x}\|^2/\delta_x^2 = r_{\min}^2.$$

On the other hand,

$$\|u_0-\widehat{u}\|_{Z_1M_1}^2 = \sigma_0/\tau_0\|x_0-\widehat{x}\|^2 - 2\sigma_0\langle x^0-\widehat{x}, [\nabla K(x^0)]^*(y^0-\widehat{y})\rangle + \|y_0-\widehat{y}\|^2,$$

therefore, using Cauchy's inequality and estimating $\|\nabla K(x^0)\| \le R_K$, we arrive at

$$r_{x,0}^2 \le (2\mu\|x_0 - \widehat{x}\|^2 + (1 + \sigma_0\tau_0 R_K^2)\|y_0 - \widehat{y}\|^2)/(\delta\mu\delta_x^2).$$

Similar to the derivations of (4.3) or (4.5), from (4.2), we obtain $\sigma_0\tau_0 R_K^2 \le 1 - \kappa \le 1$ which leads to $r_{x,0}^2 \le r_{\max}^2$. Summarizing the estimates derived in this paragraph, for (4.8) it is enough to show $\delta_y^2 r_y^2 \ge \delta_x^2 r_{\max}^2 \mu(1-\delta)\delta/(\kappa-\delta)$, which follows from the assumed bound on r_y since $\delta, \delta_x \in [0,1)$. Therefore, $\mathcal{U}_i \subseteq \mathbb{B}(\widehat{x}, \delta_x r_{x,i}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$.

For the basis of induction, since (4.2) holds, we can apply Lemma 3.2 to $\|u_0 - \widehat{u}\|_{Z_1 M_1}$ to verify $u_0 \in \mathcal{U}_0 \subseteq \mathbb{B}(\widehat{x}, \delta_x r_{x,0}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$. Moreover, since $\sigma_1 = \sigma_0/(\omega_0(1 + 2\sigma_0\widetilde{\gamma}_{F^*})) \le \sigma_0/\omega_0 = \mu\tau_0/\omega_0$, the bound (4.7) for $i = 0$ follows from (4.6) and the derived bounds $r_{\min} \le r_{x,0} \le r_{\max}$. Therefore, the basis of the induction holds.

For the inductive step, suppose $u^N \in \mathbb{B}(\widehat{x}, \delta_x r_{x,N}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$ and (4.7) holds for $i = N$. We can apply Lemma 3.7 to obtain $u^{N+1} \in \mathbb{B}(\widehat{x}, r_{x,N}) \times \mathbb{B}(\widehat{y}, r_y)$ and $\bar{x}^{N+1} \in \mathbb{B}(\widehat{x}, r_{x,N})$. From (4.2) follows $\phi_{N+1} \ge \phi_N$ and therefore $r_{x,N+1} \le r_{x,N} \le r_{\max}$ as well as $\mathbb{B}(\widehat{x}, r_{x,N}) \times \mathbb{B}(\widehat{y}, r_y) \subseteq \mathbb{B}(\widehat{x}, r_{\max}) \times \mathbb{B}(\widehat{y}, r_y) \subseteq \mathcal{U}(\rho_x, \rho_y)$. Consequently, $u_{N+1} \in \mathcal{U}(\rho_x, \rho_y)$ and $\bar{x}^{N+1} \in \mathcal{X}_K$.

Then using Theorems 2.1 and 3.6, (CI) is satisfied for $i \le N$ with $\Delta_{N+1} \le 0$, which after using Lemma 3.2, turns into

$$\delta\phi_{N+1}\|x^{N+1} - \widehat{x}\|^2 + \psi_{N+2}\tfrac{\kappa-\delta}{1-\delta}\|y^{N+1} - \widehat{y}\|^2 \le \|u^0 - \widehat{u}\|_{Z_1 M_1}^2.$$

In other words, $u^{N+1} \in \mathcal{U}_{N+1} \subseteq \mathbb{B}(\widehat{x}, \delta_x r_{x,N+1}) \times \mathbb{B}(\widehat{y}, \delta_y r_y)$.

From (4.2) we deduce $\tau_{N+1} = \eta_{N+1}/\phi_{N+1} = \tau_N/(\omega_N(1+2\tau_N\widetilde{\gamma}_G))$. Similarly, $\sigma_{N+1} = \sigma_N/(\omega_N(1+2\sigma_N\widetilde{\gamma}_{F^*}))$ and $r_{x,N+1} = r_{x,N}/\sqrt{1 + 2\tau_N\widetilde{\gamma}_G}$. Consequently, using $\omega_i\sqrt{1 + 2\tau_i\widetilde{\gamma}_G} \ge 1$ and $r_{x,N+1} \le r_{x,N}$, it follows that

$$\frac{\tau_{N+1}}{r_{x,N+1}} = \frac{1}{\omega_i\sqrt{1 + 2\tau_N\widetilde{\gamma}_G}}\frac{\tau_N}{r_{x,N}} \le \frac{1-\delta_x}{2R_K r_y + 2L\|P_{\mathrm{NL}}\widehat{y}\| r_{x,N+1}}, \quad \text{and} \tag{4.9}$$

$$\sigma_{N+2} r_{x,N+1} = \frac{\sigma_{N+1} r_{x,N}}{\omega_i\sqrt{1 + 2\tau_N\widetilde{\gamma}_G}(1 + 2\sigma_N\widetilde{\gamma}_{F^*})} \le \frac{2(1-\delta_y) r_y}{L r_{x,N+1} + 2R_K}. \tag{4.10}$$

This completes the induction. Then Assumption 4.1(v) holds since in the induction step we showed $u_{N+1} \in \mathcal{U}(\rho_x, \rho_y)$ and $\bar{x}^{i+1} \in \mathcal{X}_K$. □

Remark 4.2. *The condition $\psi_1 = 1$ is without loss of generality, as we can always rescale all the testing variables ψ_i and ϕ_i by a constant in (4.2).*

Corollary 4.3. *The claims of Lemma 4.1 are valid for any $0 < \omega_i \le 1$ if the step lengths $\tau_i \equiv \tau$ and $\sigma_i \equiv \sigma$ are constant.*

Proof. Since the step lengths are chosen constant in (4.2), it is no longer necessary to update $r_{x,i}$ and, consequently, verify (4.7) in the induction. The remaining steps follow those in the proof of Lemma 4.1. □

Corollary 4.4. *Assume Assumptions 3.1 and 3.3 hold for $\mathcal{X}_K = \mathcal{X}_G = X$ and $R_K := \sup_{x\in X}\|\nabla K(x)\| < \infty$. Then Assumption 4.1 holds for any large enough $\rho_x, \rho_y > 0$ that $\operatorname{dom} F^* \subseteq \mathbb{B}_{\mathrm{NL}}(\widehat{y}, \rho_y)$.*

Proof. The result follows immediately from the assumptions and Remark 3.8. □

4.3 DUAL LOCALITY

We now refine the choices that led to the bounds (4.3).

Theorem 4.5 (Convergence without rates). *Suppose Assumption 4.1 holds. Choose step lengths $\tau_i \equiv \tau$, $\sigma_i \equiv \sigma$, and $\omega_i \equiv 1$. Assume $\gamma_G + \theta > L\rho_y$, that $\nabla K(x^i)x \to \nabla K(x^*)x$ if $x^i \rightharpoonup x^*$ for all $x \in X$, and for some $0 \le \delta \le \kappa < 1$ the bounds*

$$\tau < \frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2(\gamma_G+\theta)-L\rho_y} + 3\rho_y\right)}, \quad \textit{and} \quad \sigma\tau < \frac{1-\kappa}{R_K^2}. \tag{4.11}$$

If either $H(u)$ is maximal monotone in $\mathcal{X} \times \mathcal{Y}$ or $(x,y) \mapsto ([\nabla K(x)]^* y, K(x))$ is weak-to-strong continuous in $\mathcal{X}\times\mathcal{Y}$, then the sequence $\{u^i\}$ converges weakly to some $u^* \in H^{-1}(0)$, possibly different from $\widehat{u}$. If $(x,y) \mapsto ([\nabla K(x)]^* y, K(x))$ is only weak-to-weak continuous, but Assumption 4.1(iii) and (iv) hold at any weak limit $u^* = (x^*, y^*)$ of $\{u^i\}$ in addition to $\widehat{u}$, then the sequence of u^i converges strongly to some $u^* \in H^{-1}(0)$.

Proof. We recall that (4.2) implies (3.5), (3.14). By taking $\widetilde{\gamma}_G = \widetilde{\gamma}_{F^*} = 0$, and any constants ψ and ϕ such that $\psi\sigma = \phi\tau$, we verify (4.2a) and (4.2b). We take $\alpha_1 = \alpha_2 = 0$. Then the second part of (4.2c) holds. Since $\omega_i \equiv 1$, (4.2d) holds with maximal $\zeta := \gamma_G + \theta - L\rho_y/2$. With the selected ζ and ω_i, (4.11) is equivalent to (4.3); therefore the first part of (4.2c) and (4.2e) hold. Hence (4.2) holds.

We will now apply Proposition 2.2. Of its assumptions, (CI) and the self-adjointness of $Z_{i+1}M_{i+1}$ are verified by the combination of Theorem 3.6 and Lemma 3.2, the requirements of which immediately follow from (4.2) shown and Assumption 4.1. In fact, since the bounds (4.11) are strict, Theorem 3.6 holds with $\Delta_{i+1} \le -\hat{\delta}\|u^{i+1} - u^i\|^2$ for some $\hat{\delta} > 0$. Combining (3.7) and (4.11), we verify Proposition 2.2(i). Then (iii) follows from the assumed constant step lengths and the assumption that $\nabla K(x^i)x \to \nabla K(x^*)x$ if $x^i \rightharpoonup x^*$.

It only remains to show the condition (ii) of Proposition 2.2. If $H(u)$ is maximal monotone, the necessary inclusion follows from the fact that maximal monotone operators have sequentially weakly–strongly closed graphs [3, Proposition 20.38]. Otherwise, for any $x^{i+1} \rightharpoonup x^*$ and $y^{i+1} \rightharpoonup y^*$ we have $W_{i+1} \equiv W$ and

$$v_{i+1} := W \begin{pmatrix} -[\nabla K(x^{i+1})]^* y^{i+1} \\ K(x^{i+1}) \end{pmatrix} + V_{i+1}(u^{i+1}) \in W \begin{pmatrix} \partial G(x^{i+1}) \\ \partial F^*(y^{i+1}) \end{pmatrix} := A(u^{i+1}). \tag{4.12}$$

We need to show that $v_{i+1} \rightharpoonup v^* := (-[\nabla K(x^*)]^* y^*, K(x^*))$ and $v^* \in A(u^*)$, which is tantamount to the inclusion $u^* \in H^{-1}(0)$. Since $Z_{i+1}M_{i+1}(u^{i+1} - u^i) \to 0$, it follows that $V_{i+1}(u^{i+1}) \to 0$ from the definition of V_{i+1} in (2.2) and (2.3).

If $[\nabla K(x)]^* y$ and $K(x)$ are weak-to-strong continuous, $v_{i+1} \to v^*$ and the required inclusion $v^* \in A(u^*)$ follows from the fact that, in the case of convex lower semicontinuous functions, the graph of a subgradient mapping (A in our case) is sequentially weakly–strongly closed ([3, Proposition 16.36]). Therefore, $u^i \rightharpoonup u^* \in H^{-1}(0)$.

If $[\nabla K(x)]^* y$ and $K(x)$ are only weak-to-weak continuous and Assumption 4.1(iii) and (iv) hold at u^*, then $v_{i+1} \rightharpoonup v^*$. We apply [3, Corollary 20.59 (iii)], which states that if A is maximally monotone, $(u_i, v_i) \rightharpoonup (u^*, v^*)$ with $v_i \in A(u_i)$, and $\varlimsup_{i\to\infty} \langle u_i - u^*, v_i - v^* \rangle \le 0$, then

$\underline{(u_i, v_i)} \to (u^*, v^*)$ and $v^* \in \underline{A(u^*)}$. In our case, A is as in (4.12), and $V_{i+1}(u^{i+1}) \to 0$. Consequently $\overline{\lim}_{i\to\infty} \langle u_i - u^*, v_i - v^*\rangle = \overline{\lim}_{i\to\infty}\ q_i$ for

$$q_i := \langle [\nabla K(x^*)]^* y^* - [\nabla K(x^{i+1})]^* y^{i+1}, x^{i+1} - x^*\rangle + \langle K(x^{i+1}) - K(x^*), y^{i+1} - y^*\rangle.$$

Note that $\|y^{i+1} - y^*\|_{P_{\mathrm{NL}}} \le 2\rho_y$ because $\|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}}, \|\widehat{y} - y^*\|_{P_{\mathrm{NL}}} \le \rho_y$. With this, (3.2), and both Assumption 4.1(iii) and (iv) at u^*, we bound

$$\begin{aligned} q_i &= \langle K(x^{i+1}) - K(x^*) + \nabla K(x^{i+1})(x^* - x^{i+1}), y^{i+1} - y^*\rangle \\ &\quad - \langle (\nabla K(x^{i+1}) - \nabla K(x^*))(x^{i+1} - x^*), y^*\rangle \le \left(L\rho_y - \theta\right) \|x^{i+1} - x^*\|^2. \end{aligned}$$

Since $\gamma_G + \theta > L\rho_y$, this proves $q_i \le 0$ if $\gamma_G = 0$. If $\gamma_G > 0$, we can apply the argument to $A - \left(\begin{smallmatrix}\gamma_G I & 0\\ 0 & 0\end{smallmatrix}\right)$, which is monotone. Therefore, the conditions of [3, Corollary 20.59 (iii)] hold. Consequently, Proposition 2.2(ii) holds with $u^i \to u^* \in H^{-1}(0)$ strongly. ◻

We can choose different step lengths on the "linear" and "nonlinear" dual subspaces:

Corollary 4.6. *Let us write*

$$\nabla K(x)(\Delta x) := P_{\mathrm{L}} K_{\mathrm{L}} \Delta x + P_{\mathrm{NL}} \nabla K_{\mathrm{NL}}(x)(\Delta x) \quad \textit{and} \quad R_{\mathrm{NL}} := \sup_{x\in X} \|\nabla K_{\mathrm{NL}}(x)\|. \tag{4.13a}$$

If we choose distinct step lengths on the subspaces Y_{L} and Y_{NL} as

$$\Sigma_i := \sigma_{i,\mathrm{L}} P_{\mathrm{L}} + \sigma_{i,\mathrm{NL}} P_{\mathrm{NL}} \quad \textit{and} \quad \Psi_i := \psi_{i,\mathrm{L}} P_{\mathrm{L}} + \psi_{i,\mathrm{NL}} P_{\mathrm{NL}}, \tag{4.13b}$$

then the claims of Theorem 4.5 continue to hold if we replace (4.11) by

$$\tau < \frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2(\gamma_G+\theta) - L\rho_y} + 3\rho_y\right)}, \quad \textit{and} \quad \begin{cases} \sigma_{\mathrm{L}}\tau < \frac{1-\kappa}{\|K_{\mathrm{L}}\|^2}, \\ \sigma_{\mathrm{NL}}\tau < \frac{1-\kappa}{R_{\mathrm{NL}}^2}. \end{cases}$$

Proof. The proof repeats that of Theorem 4.5, but now with (3.14) leading to two variants of the last condition in (4.2a) for $\psi_{i,\mathrm{L}}$ and $\psi_{i,\mathrm{NL}}$. ◻

To ensure weak convergence above, we had to impose additional conditions on K. These will not be required if G and F^* are regular enough to give convergence rates.

If $\gamma_{F^*} = 0$, we need to take $\widetilde{\gamma}_{F^*} = \alpha_1 = \alpha_2 = 0$ to satisfy the second part of (4.2c). With $\widetilde{\gamma}_{F^*} = 0$, the second part of (4.2b) forces $\psi_i \equiv \psi_0$. Then (4.2a) holds when $\omega_i = \sigma_i/\sigma_{i+1}$ and $\eta_i := \psi_0\sigma_i = \phi_i\tau_i$. Taking into account (4.3), we would like to maintain $\sigma_i\tau_i \equiv c_0$ for a constant c_0. Therefore $\phi_i = c_0\psi_0/\tau_i^2$. If now $\widetilde{\gamma}_G > 0$, we obtain from (4.2b) the update rule

$$\tau_{i+1} = \tau_i\omega_i, \quad \sigma_{i+1} = \sigma_i/\omega_i, \quad \omega_i = 1/\sqrt{1 + 2\tau_i\widetilde{\gamma}_G}. \tag{4.14}$$

As shown in [7] and [25, Remark 3.2], this update rules causes τ_N to go to zero at the rate $O(1/N)$, hence, ϕ_i to grow at the rate $\Omega(N^2)$. Thus we obtain:

Theorem 4.7 (Acceleration). *Suppose Assumption 4.1 holds. Let* $\widetilde{\gamma}_G = (\gamma_G+\theta)(1-\delta/6)-L\rho_y/2-\zeta > 0$ *for some* $0 < \delta \le \kappa < 1$, $\zeta > 0$, *and apply the update rules (4.14) for initial iterates satisfying*

$$\tau_0 \le \frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2\zeta} + 3\rho_y\right)}, \quad \text{and} \quad \tau_0\sigma_0 \le \frac{1-\kappa}{R_K^2}. \tag{4.15}$$

Then $\|x^i - \widehat{x}\|^2$ *converges to zero at the rate* $O(1/N^2)$.

Proof. The first stages of the proof are similar to Theorem 4.5: We have to verify (4.2); but then we do not need to verify the conditions of Proposition 2.2, as we directly use Theorem 2.1 and afterwards estimate the convergence rate from $Z_{N+1}M_{N+1}$.

We start with (4.2). The discussion above (4.14) and the second bound of (4.15) verify (4.2a)–(4.2c). Using (4.14) and our choice of $\widetilde{\gamma}_G$, we estimate

$$1/\omega_i \le 1/\omega_0 \le \sqrt{1 + 2\delta(\gamma_G + \theta)/(3L\rho_y)} \le 1 + \delta(\gamma_G + \theta)/(3L\rho_y).$$

This quickly shows (4.2d) with $\alpha_2 = 0$. The remaining (4.2e) follows via (4.3) from the first bound of (4.15).

We now apply Theorems 2.1 and 3.6 to arrive at (DI) with each $\Delta_{i+1} \le 0$. Then, using Lemma 3.2, we conclude

$$\delta\phi_N\|x^N - \widehat{x}\|^2 \le \|u^0 - \widehat{u}\|^2_{Z_1M_1} \tag{4.16}$$

and obtain the desired convergence rate due to ϕ_N growing as $\Omega(N^2)$. ◻

Remark 4.8. *The update rule (4.14) on* ω_i *is consistent with the bound required in Lemma 4.1. Consequently, if for the starting point* u^0 *and* $\mathcal{U}(\rho_x, \rho_y)$, *the conditions of Lemma 4.1, including the initialization bounds (4.6) on* $\tau_0, \sigma_0, \omega_0$, *and* u^0, *are satisfied, all the iterations* $\{u^i\}_{i\in\mathbb{N}}$ *will belong to* $\mathcal{U}(\rho_x, \rho_y)$ *and verify Assumption 4.1(v).*

Corollary 4.9. *With the split steps (4.13) on* Y_{L} *and* Y_{NL}, *the claims of Theorem 4.7 hold if the rules for* σ_{i+1} *and* σ_0 *in (4.14) and (4.15) are replaced by*

$$\left.\begin{aligned} \sigma_{i+1,\mathrm{L}} &= \sigma_{i,\mathrm{L}}/\omega_i, \\ \sigma_{i+1,\mathrm{NL}} &= \sigma_{i,\mathrm{NL}}/\omega_i, \end{aligned}\right\} \quad \text{and} \quad \left\{\begin{aligned} \tau_0\sigma_{0,\mathrm{L}} &\le \tfrac{1-\kappa}{\|K_{\mathrm{L}}\|^2}, \\ \tau_0\sigma_{0,\mathrm{NL}} &\le \tfrac{1-\kappa}{R_{\mathrm{NL}}^2}. \end{aligned}\right. \tag{4.17}$$

Finally, if ∂F^* is strongly monotone as well, an algorithm with constant step lengths converges linearly in the primal variable according to the following theorem.

Theorem 4.10 (Linear convergence). *Suppose Assumption 4.1 holds. Let* $\widetilde{\gamma}_{F^*} = \gamma_{F^*} > 0$, *and* $\widetilde{\gamma}_G = (\gamma_G + \theta)(1 - \delta/3) - L\rho_y/2 - \zeta > 0$ *for some small* $\zeta > 0$. *Take* $0 \le \delta \le \kappa < 1$, *and* $\tau_i \equiv \tau$, $\sigma_i \equiv \sigma$, *and* $\omega_i \equiv \omega$ *for*

$$0 < \tau \le \min\left\{\frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2\zeta}+3\rho_y\right)}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{F^*}/\widetilde{\gamma}_G}}{R_K}\right\}, \quad \sigma := \frac{\widetilde{\gamma}_G}{\widetilde{\gamma}_{F^*}}\tau, \quad \text{and} \quad \omega := \frac{1}{1+2\widetilde{\gamma}_G\tau}. \tag{4.18}$$

Then $\|u^i - \widehat{u}\|^2$ *converges to zero with the rate* $O(1/(1 + 2\widetilde{\gamma}_G\tau)^N)$.

Proof. The structure of the proof is as of Theorem 4.7: We verify (4.2) and then estimate $Z_{N+1}M_{N+1}$ in (DI).

To verify (4.2), we take $\psi_0 = 1/\sigma$, and $\phi_0 = 1/\tau$. Then the σ-rule of (4.18) verifies (4.2b). The latter applied repeatedly gives

$$\phi_N\tau = \psi_N\sigma = (1 + 2\widetilde{\gamma}_G\tau)^N. \tag{4.19}$$

With $N = i$ this proves (4.2a) for $\omega_i = \omega$ given by (4.18). The second inequality of (4.2c) holds due to $\widetilde{\gamma}_{F^*} = \gamma_{F^*} > 0$ by taking $\alpha_1 = \alpha_2 = 0$. For the first inequality, from (4.18), $\tau \le \delta/(3\rho_y L)$, consequently, $1/\omega \le 1 + 2(\gamma_G + \theta)\delta/(3\rho_y L)$. This gives $\gamma_G + \theta - \widetilde{\gamma}_G - L\rho_y/(2\omega) - \zeta > 0$, as required with $\alpha_2 = 0$. It remains to prove the first inequality of (4.2c) and (4.2e), which follow via (4.3) from the bound on τ in (4.18).

Finally, we apply Lemma 3.2, (4.19), and Theorem 2.1 to conclude that

$$(1 + 2\widetilde{\gamma}_G\tau)^N\left(\frac{\delta}{2\tau}\|x^N - \widehat{x}\|^2 + \frac{\kappa - \delta}{2\sigma(1-\delta)}\|y^N - \widehat{y}\|^2\right) \le \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1M_1}.$$

This gives the desired convergence rate. ◻

Remark 4.11. *To verify Assumption 4.1(v) in this case, one can use Corollary 4.3 by checking the bounds (4.6) for the starting point u^0, selected τ, σ, ω, and $\mathcal{U}(\rho_x, \rho_y)$.*

Corollary 4.12. *With the split steps (4.13) on Y_{L} and Y_{NL}, and $\Gamma_{F^*} := \gamma_{\mathrm{L}}P_{\mathrm{L}} + \gamma_{\mathrm{NL}}P_{\mathrm{NL}}$, the claims of Theorem 4.10 continue to hold if we take $\sigma_{\mathrm{L}} = \widetilde{\gamma}_{\mathrm{L}}^{-1}\widetilde{\gamma}_G\tau$ and $\sigma_{\mathrm{NL}} = \widetilde{\gamma}_{\mathrm{NL}}^{-1}\widetilde{\gamma}_G\tau$ with $\widetilde{\gamma}_{\mathrm{L}} = \gamma_{\mathrm{L}} > 0$ and $\widetilde{\gamma}_{\mathrm{NL}} = \gamma_{\mathrm{NL}} > 0$, and replace (4.18) by*

$$\tau \le \min\left\{\frac{\delta}{L\left(\frac{L\|P_{\mathrm{NL}}\widehat{y}\|^2}{2\zeta} + 3\rho_y\right)}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{\mathrm{L}}/\widetilde{\gamma}_G}}{\|K_{\mathrm{L}}\|}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{\mathrm{NL}}/\widetilde{\gamma}_G}}{\|R_{\mathrm{NL}}\|}\right\}.$$

Remark 4.13 (Global convergence). *Following Remark 3.8 and Corollary 4.4, if Assumption 4.1 holds for $\mathcal{X}_K = \mathcal{X}_G = X$ and $R_K := \sup_{x\in X}\|\nabla K(x)\| < \infty$, then ρ_x can be taken infinitely large. Consequently, the convergence results of Theorem 4.5, Theorem 4.7, and Theorem 4.10 will hold globally provided $\operatorname{dom} F^* \subseteq \mathbb{B}_{\mathrm{NL}}(\widehat{y}, \rho_y)$.*

4.4 Primal locality; dual strong convexity

With $\alpha_1 = \alpha_2 = 1$ and the additional requirement $\gamma_{F^*} > 0$, the results of Section 4.1 can be reformulated using locality in the primal variable rather than dual, i.e. relying on ρ_x instead of ρ_y in the step length bounds sufficient for convergence. Since the main differences from the proofs of Section 4.1 are in replacing the instances of (4.3) with (4.5), the proofs only indicate those differences.

Theorem 4.14 (Convergence without rates). *Suppose Assumption 4.1 holds for $\gamma_{F^*} > 0$, $\gamma_G + \theta > 0$; $\rho_x < 4\sqrt{(\gamma_G + \theta)\gamma_{F^*}}/L$; and the step lengths $\tau_i \equiv \tau$ and $\sigma_i \equiv \sigma$. Also assume that $\nabla K(x^i)x \to \nabla K(x^*)x$ if $x^i \rightharpoonup x^*$ for all $x \in X$, and the step lengths satisfy for some $0 \le \delta \le \kappa < 1$ the bounds*

$$\tau < \frac{\delta(16(\gamma_G + \theta)\gamma_{F^*}/L^2 - \rho_x^2)}{8\|P_{\mathrm{NL}}\widehat{y}\|^2\gamma_{F^*} + 12L\|P_{\mathrm{NL}}\widehat{y}\|\rho_x^2 + 72(\gamma_G + \theta)\rho_x^2}, \quad \text{and} \quad \sigma\tau < \frac{1-\kappa}{R_K^2}. \tag{4.20}$$

If either $H(u)$ is maximal monotone in $\mathcal{X} \times \mathcal{Y}$ or $(x, y) \mapsto ([\nabla K(x)]^* y, K(x))$ is weak-to-strong continuous in $\mathcal{X} \times \mathcal{Y}$, then the sequence $\{u^i\}$ converges weakly to some $u^* \in H^{-1}(0)$, possibly different from $\widehat{u}$. If instead $(x, y) \mapsto ([\nabla K(x)]^* y, K(x))$ is only weak-to-weak continuous, but

$$\langle (\nabla K(x^i) - \nabla K(x^*))(x^i - x^*), y^* \rangle \geq \theta^* \|x^i - x^*\|^2 \tag{4.21}$$

for any weak limit $u^* = (x^*, y^*)$ of $\{u^i\}$ and some $\theta^* \in \mathbb{R}$ such that $\gamma_G + \theta^* \geq L\rho_y$, then the sequence of u^i converges strongly to some $u^* \in H^{-1}(0)$.

Proof. We take $\alpha_1 = \alpha_2 = 1$ in (4.2). Since $\rho_x < 4\sqrt{(\gamma_G + \theta)\gamma_{F^*}}/L$, the bound (4.4) follows with $\omega_i = 1$ and $\widetilde{\gamma}_G = \widetilde{\gamma}_{F^*} = 0$. Then (4.2d) are satisfied for the choices $\beta_1 = L\rho_x\beta_2/(\gamma_{F^*}\beta_2 - L\rho_x)$, $\zeta = \gamma_G + \theta - L\beta_2\rho_x/16$, and β_2 leading to (4.5). The latter is equivalent to (4.20) for the selected $\widetilde{\gamma}_G, \widetilde{\gamma}_{F^*}$, and ω. And from the derivation of (4.5), we get bounds (4.2c) and (4.2e). The remaining steps of the proof repeat those of Theorem 4.5. □

Corollary 4.15. *With the split steps (4.13), the claims of Theorem 4.14 continue to hold if all the instances of γ_{F^*} in the formulation of Theorem 4.14 are replaced with γ_{NL}, and the σ-rule of (4.20) is split into $\sigma_{\mathrm{L}}\tau < (1-\kappa)/\|K_{\mathrm{L}}\|^2$ and $\sigma_{\mathrm{NL}}\tau < (1-\kappa)/R_{\mathrm{NL}}^2$.*

Theorem 4.16 (Acceleration). *Suppose Assumption 4.1 holds for $\gamma_{F^*} > 0$, $\widetilde{\gamma}_G = \gamma_G + \theta - \zeta > 0$ for some small $\zeta > 0$, and $\rho_x < 4\sqrt{\zeta\gamma_{F^*}}/(L\sqrt{1 + \tau_0\widetilde{\gamma}_G})$. Apply the update rules (4.14), assuming for some $0 < \delta \leq \kappa < 1$ the initialization conditions*

$$\tau_0 \leq \frac{\delta(16\zeta\gamma_{F^*}/L^2 - \rho_x^2)}{8\|P_{\mathrm{NL}}\widehat{y}\|^2\gamma_{F^*} + 12L\|P_{\mathrm{NL}}\widehat{y}\|\rho_x^2 + 72\zeta\rho_x^2}, \quad \tau_0\sigma_0 \leq \frac{1-\kappa}{R_K^2}. \tag{4.22}$$

Then $\|x^i - \widehat{x}\|^2$ converges to zero at the rate $O(1/N^2)$.

Proof. Similar to Theorem 4.14, the only difference with the proof of Theorem 4.7 is in the verification of (4.2d). With $\alpha_1 = \alpha_2 = 1$ in (4.2), $\omega_i \geq \omega_0 \geq 1/(1 + \tau_0\widetilde{\gamma}_G)$, consequently, (4.4) holds if $\rho_x < 4\sqrt{\zeta\gamma_{F^*}}/(L\sqrt{1 + \tau_0\widetilde{\gamma}_G})$, as was assumed. The remaining steps follow the proof of Theorem 4.7. □

Corollary 4.17. *With the split steps of (4.13), the claims of Theorem 4.16 continue to hold if all the instances of γ_{F^*} in the formulation of Theorem 4.16 are replaced with γ_{NL}, the σ update rule of (4.14) is replaced with (4.17) and (4.22) initialization is split into $\sigma_{0,\mathrm{L}}\tau < (1-\kappa)/\|K_{\mathrm{L}}\|^2$ and $\sigma_{0,\mathrm{lNL}}\tau < (1-\kappa)/R_{\mathrm{NL}}^2$.*

Theorem 4.18 (Linear convergence). *Suppose Assumption 4.1 holds for $\widetilde{\gamma}_G = \gamma_G + \theta - \zeta_1 > 0$ and $\widetilde{\gamma}_{F^*} = \gamma_{F^*} - \zeta_2 > 0$ for some small $\zeta_1, \zeta_2 > 0$, and $\rho_x < 4\sqrt{\zeta_1\zeta_2}/(L\sqrt{1 + 2\tau\widetilde{\gamma}_G})$. Pick $0 \leq \delta \leq \kappa < 1$, and take constant $\omega_i \equiv \omega = 1/(1 + 2\tau_0\widetilde{\gamma}_G)$ and step lengths $\tau_i \equiv \tau$, $\sigma_i \equiv \sigma$, satisfying*

$$\tau \leq \min\left\{\frac{\delta(16\zeta_1\zeta_2/L^2 - \rho_x^2)}{8\|P_{\mathrm{NL}}\widehat{y}\|^2\zeta_2 + 12L\|P_{\mathrm{NL}}\widehat{y}\|\rho_x^2 + 72\zeta_1\rho_x^2}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{F^*}/\widetilde{\gamma}_G}}{R_K}\right\}, \quad \text{and} \quad \sigma = \frac{\widetilde{\gamma}_G}{\widetilde{\gamma}_{F^*}}\tau. \tag{4.23}$$

Then $\|u^i - \widehat{u}\|^2$ converges to zero with the rate $O(1/(1 + 2\widetilde{\gamma}_G\tau)^N)$.

Proof. Similar to Theorem 4.14, now with the nonzero $\widetilde{\gamma}_G$ and $\widetilde{\gamma}_{F^*}$; the only difference with the proof of Theorem 4.18 is in the verification of (4.2d). With $\alpha_1 = \alpha_2 = 1$ in (4.2), $\omega_i \equiv 1/(1+2\tau\widetilde{\gamma}_G)$, consequently, the bound (4.4) holds if $\rho_x \le 4\sqrt{\zeta\gamma_{F^*}}/(L\sqrt{1+2\tau\widetilde{\gamma}_G})$, as was assumed. The remaining steps follow the proof of Theorem 4.10. ◻

Corollary 4.19. *With the split steps of* (4.13)*, the claims of Theorem 4.10 continue to hold if we take $\sigma_{\mathrm{L}} = \widetilde{\gamma}_{\mathrm{L}}^{-1}\widetilde{\gamma}_G\tau$ and $\sigma_{\mathrm{NL}} = \widetilde{\gamma}_{\mathrm{NL}}^{-1}\widetilde{\gamma}_G\tau$ with $\widetilde{\gamma}_{\mathrm{L}} = \gamma_{\mathrm{L}} > 0$ and $\widetilde{\gamma}_{\mathrm{NL}} = \gamma_{\mathrm{NL}} > 0$, and instead of $\widetilde{\gamma}_{F^*} = \gamma_{F^*} - \zeta_2 > 0$ and* (4.23)*, we require $\widetilde{\gamma}_{\mathrm{NL}} = \gamma_{\mathrm{NL}} - \zeta_2 > 0$ and*

$$\tau \le \min\left\{\frac{\delta(16\zeta_1\zeta_2/L^2 - \rho_x^2)}{8\|P_{\mathrm{NL}}\widehat{y}\|^2\zeta_2 + 12L\|P_{\mathrm{NL}}\widehat{y}\|\rho_x^2 + 72\zeta_1\rho_x^2}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{\mathrm{L}}/\widetilde{\gamma}_G}}{\|K_{\mathrm{L}}\|}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{\mathrm{NL}}/\widetilde{\gamma}_G}}{\|R_{\mathrm{NL}}\|}\right\}.$$

4.5 HYPOMONOTONE SADDLE TERM GRADIENT

In Assumption 4.1 we required the gradient of $\langle K(x), \widehat{y}\rangle$ to be hypomonotone at $\widehat{x}$:

$$\langle (\nabla K(x) - \nabla K(\widehat{x}))(x - \widehat{x}), \widehat{y}\rangle \ge \theta\|x - \widehat{x}\|^2 \quad (x \in \mathcal{X})$$

for some factor $\theta \in \mathbb{R}$. In the proof of Theorem 4.5 we saw that if this property holds at any weak limit of $\{u^i\}_{i\in\mathbb{N}}$ with $\theta + \gamma > 0$, the convergence with fixed step lengths becomes locally strong. In this section, we explore what improvements to the main result can be expected if the gradient of $\langle K(x), \widehat{y}\rangle$ is hypomonotone at any $x \in \mathcal{X}$.

Lemma 4.20. *In addition to the requirements of Theorem 3.6, suppose $x \mapsto \nabla K(x)^*\widehat{y}$ to be hypomonotone in $\mathcal{X}$:*

$$\langle (\nabla K(x) - \nabla K(x'))(x - x'), \widehat{y}\rangle \ge \theta\|x - x'\|^2 \quad (x, x' \in \mathcal{X}_K), \tag{4.24}$$

with the factor $\theta > -L'$ for $L' := L\|\widehat{y}\|_{P_{\mathrm{NL}}}$, and the neighborhood $\mathcal{X}_K$ convex. Then (CI) *is satisfied with*

$$\Delta_{i+1} \le \frac{\theta L'}{L' + \theta}\left(\eta_{i+1}\|x^{i+1} - \widehat{x}\|^2 - \eta_i\|x^i - \widehat{x}\|^2\right)$$

if for some $0 \le \delta \le \kappa < 1$ we have

$$\tau_i \le \frac{2\delta}{L' + \theta + 2L(\omega_i + 2)\rho_y}, \qquad \tau_i\sigma_i \le \frac{1-\kappa}{\|\nabla K(x^i)\|^2}, \tag{4.25}$$

$$0 \le \widetilde{\gamma}_{F^*} \le \gamma_{F^*}, \quad \text{and} \qquad \widetilde{\gamma}_G \le \gamma_G + \frac{1}{2\omega_i}\left(\frac{2\theta L'}{\theta + L'} - L\rho_y\right). \tag{4.26}$$

Proof. We abbreviate $A(x) := \nabla K(x)^*\widehat{y}$. Then A is hypomonotone with factor θ, and Lipschitz with factor $L' := L\|\widehat{y}\|_{P_{\mathrm{NL}}}$. We begin by showing that the Cauchy inequality that introduced ζ into (3.10) is no longer needed to estimate the nonlinear preconditioner V'_{i+1} in Lemma 3.4. Indeed, observe that the map $Q(x) := \langle K(x), \widehat{y}\rangle - \theta\frac{1}{2}\|x\|^2$ is convex within $\mathcal{X}_K$, as by (4.24) its differential $\nabla Q(x) = A(x) - \theta x$ is monotone in this convex domain. Moreover, ∇Q is Lipschitz with the constant $L' - \theta$. Indeed, with (3.2) we estimate

$$\begin{aligned} Q(x') - Q(x) - \nabla Q(x)(x' - x) &\le \frac{\theta}{2}(\|x\|^2 - \|x'\|^2) + \theta\langle x, x' - x\rangle + \frac{L'}{2}\|x - x'\|^2 \\ &= \frac{L' - \theta}{2}\|x - x'\|^2. \end{aligned}$$

If $\theta > 0$, (3.1a) and (4.24) establish $L' > \theta$. Therefore, always $L' - \theta > 0$. Applying [3, Theorem 18.15], we conclude that ∇Q is Lipschitz with the constant $L' - \theta > 0$.

Hence, ∇Q is also $(L' - \theta)^{-1}$-cocoercive, see, e.g., [3, Theorem 18.1], so that:

$$
\begin{aligned}
0 &= \frac{L'-\theta}{L'+\theta}\left(\langle \nabla Q(x^i) - \nabla Q(\widehat{x}), x^i - \widehat{x}\rangle + \theta\|x^i - \widehat{x}\|^2 - \langle A(x^i) - A(\widehat{x}), x^i - \widehat{x}\rangle\right) \\
&\geq \frac{1}{L'+\theta}\|\nabla Q(x^i) - \nabla Q(\widehat{x})\|^2 + \frac{L'-\theta}{L'+\theta}\left(\theta\|x^i - \widehat{x}\|^2 - \langle A(x^i) - A(\widehat{x}), x^i - \widehat{x}\rangle\right) \\
&= \frac{1}{L'+\theta}\|A(x^i) - A(\widehat{x})\|^2 - \langle A(x^i) - A(\widehat{x}), x^i - \widehat{x}\rangle + \frac{\theta L'}{L'+\theta}\|x^i - \widehat{x}\|^2.
\end{aligned}
\tag{4.27}
$$

Next, we decompose

$$
\begin{aligned}
\langle A(x^i) - A(x^{i+1}), x^{i+1} - \widehat{x}\rangle = \langle A(x^i) - A(\widehat{x}), x^i - \widehat{x}\rangle + \langle A(x^i) - A(\widehat{x}), x^{i+1} - x^i\rangle \\
- \langle A(x^{i+1}) - A(\widehat{x}), x^{i+1} - \widehat{x}\rangle.
\end{aligned}
$$

First using Cauchy's inequality, and then (4.27), we therefore estimate

$$
\begin{aligned}
\langle A(x^i) - A(x^{i+1}), x^{i+1} - \widehat{x}\rangle &\geq \langle A(x^i) - A(\widehat{x}), x^i - \widehat{x}\rangle \frac{1}{L'+\theta}\|([\nabla K(x^i)]^* - [\nabla K(\widehat{x})]^*)\widehat{y}\|^2 \\
&\qquad - \frac{L'+\theta}{4}\|x^{i+1} - x^i\|^2 - \langle A(x^{i+1}) - A(\widehat{x}), x^{i+1} - \widehat{x}\rangle \\
&\geq \frac{\theta L'}{L'+\theta}\|x^i - \widehat{x}\|^2 - \frac{L'+\theta}{4}\|x^{i+1} - x^i\|^2 - \langle A(x^{i+1}) - A(\widehat{x}), x^{i+1} - \widehat{x}\rangle.
\end{aligned}
$$

Expanding $A(x) := \nabla K(x)^*\widehat{y}$ and using this estimate in (3.10), we obtain

$$
\begin{aligned}
D \geq \|u^{i+1} - \widehat{u}\|^2_{Q_{i+1}(0,0)} - \eta_i\left(\frac{L'+\theta}{4} + L\left(\frac{\omega_i}{2} + 1\right)\|y^{i+1} - \widehat{y}\|_{P_{\mathrm{NL}}}\right)\|x^{i+1} - x^i\|^2 \\
+ \eta_i\frac{\theta L'}{L'+\theta}\|x^i - \widehat{x}\|^2 - \eta_i\langle[\nabla K(x^{i+1}) - \nabla K(\widehat{x})](x^{i+1} - \widehat{x}), \widehat{y}\rangle.
\end{aligned}
\tag{4.28}
$$

Notice that the last term $\eta_i\langle[\nabla K(x^{i+1}) - \nabla K(\widehat{x})](x^{i+1} - \widehat{x}), \widehat{y}\rangle$ cancels out with the corresponding term in Lemma 3.5, i.e. in (3.12). Following the logic of Theorem 3.6 and rearranging some terms, (CI) is thus satisfied if

$$
\frac{1}{2}\|u^{i+1} - u^i\|^2_{\widetilde{S}_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{\widetilde{\hat{S}}_{i+1}} - \widetilde{\Delta}_{i+1} \geq -\Delta_{i+1},
$$

where for some $0 \leq \delta \leq \kappa < 1$ now

$$
\begin{aligned}
\widetilde{S}_{i+1} &:= \begin{pmatrix} \delta\Phi_i - \eta_i\left(\frac{L'+\theta}{2} + L(\omega_i + 2)\rho_y\right)I & 0 \\ 0 & \Psi_{i+1} - \frac{\eta_i^2}{1-\kappa}\nabla K(x^i)\Phi_i^{-1}[\nabla K(x^i)]^* \end{pmatrix}, \\
\widetilde{\hat{S}}_{i+1} &:= \begin{pmatrix} \eta_i\left[\Gamma_G - \widetilde{\Gamma}_G + \left(\frac{\theta L'}{(L'+\theta)\omega_i} - \frac{L\rho_y}{2\omega_i}\right)I\right] & 0 \\ 0 & \eta_{i+1}[\Gamma_{F^*} - \widetilde{\Gamma}_{F^*}] \end{pmatrix}, \quad \text{and} \\
\widetilde{\Delta}_{i+1} &:= \frac{\theta L'}{L'+\theta}\left(\eta_{i+1}\|x^{i+1} - \widehat{x}\|^2 - \eta_i\|x^i - \widehat{x}\|^2\right).
\end{aligned}
\tag{4.29}
$$

But (4.25) and (4.26) show $\frac{1}{2}\|u^{i+1} - u^i\|^2_{\widetilde{S}_{i+1}} + \|u^{i+1} - \widehat{u}\|^2_{\widetilde{\hat{S}}_{i+1}} \geq 0$, which yields the claim. □

Remark 4.21. *If Lemma 4.20 holds, we will have $\Delta_{i+1} \le \widetilde{\Delta}_{i+1}$, where $\widetilde{\Delta}_{i+1}$ is given by (4.29). However, $\sum_{i=0}^{N-1} \widetilde{\Delta}_{i+1} = \theta L'(\eta_N \|x_N - \widehat{x}\|^2 - \eta_0 \|x_0 - \widehat{x}\|^2)/(L' + \theta)$. Therefore, after application of Theorem 2.1 and Lemma 3.2, (DI) becomes*

$$\frac{1}{2}\|x^N - \widehat{x}\|^2_{A_N} + \frac{\kappa - \delta}{2(1-\delta)}\|y^N - \widehat{y}\|^2_{\psi_{N+1} I} \le \frac{1}{2}\|u^0 - \widehat{u}\|^2_{Z_1 M_1} - \eta_0 \frac{\theta L'}{L' + \theta}\|x^0 - \widehat{x}\|^2 + D_N, \tag{4.30}$$

where

$$A_N := \phi_N \left(\delta - \frac{2\theta L'}{L' + \theta}\tau_N\right) I, \quad \text{and} \quad D_N := \sum_{i=0}^{N-1} (\Delta_{i+1} - \widetilde{\Delta}_{i+1}) \le 0.$$

Hence, the convergence rate will again correspond to ϕ_N and ψ_{N+1} as long as $\tau_N \le \delta(L'+\theta)/(2\theta L') - \varepsilon$, which in fact will hold if the first inequality of (4.25) is either strict or holds with $\rho_y > 0$ or with $\theta < L'$.

In particular, for constant step lengths, Proposition 2.2 is still applicable even though now $\Delta_{i+1} - \widetilde{\Delta}_{i+1} \le -\frac{\hat{\delta}}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1} M_{i+1}}$. Indeed, if $\Delta_{i+1} - \widetilde{\Delta}_{i+1} \le -\frac{\hat{\delta}}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1} M_{i+1}}$, inequality (4.30) will still result in $i \mapsto \|u^i - \widehat{u}\|^2_{Z_{i+1} M_{i+1}}$ being nonincreasing and $\sum_{i=0}^{\infty} \frac{\hat{\delta}}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1} M_{i+1}} < \infty$.

Given Lemma 4.20 and Remark 4.21, the following corollaries immediately follow:

Corollary 4.22 (Convergence without rates: hypomonotone case). *Suppose the conditions of Theorem 4.5 are satisfied with (4.11) replaced by*

$$\tau \le \frac{2\delta}{L' + \theta + 6L\rho_y}, \quad \text{and} \quad \tau\sigma \le \frac{1-\kappa}{R_K^2}.$$

If $x \mapsto \nabla K(x)^* \widehat{y}$ is hypomonotone in X with the parameter $\theta > -L'$ and $\gamma_G + \theta L'/(L' + \theta) > L\rho_y$, the results of Theorem 4.5 still hold.

Corollary 4.23 (Acceleration: hypomonotone case). *Suppose the conditions of Theorem 4.7 are satisfied with $\widetilde{\gamma}_G = (1 - \delta/6)(\gamma_G + \theta L'/(L' + \theta)) - L\rho_y/2 > 0$, and the initialization conditions (4.15) replaced with*

$$\tau_0 < \frac{2\delta}{L' + \theta + 6L\rho_y}, \quad \text{and} \quad \tau_0\sigma_0 \le \frac{1-\kappa}{R_K^2}.$$

If $x \mapsto \nabla K(x)^* \widehat{y}$ is hypomonotone in X with the parameter $\theta > -L'$, then $\|x^i - \widehat{x}\|^2$ converges to zero at the rate $O(1/N^2)$.

Corollary 4.24 (Linear convergence: hypomonotone case). *Suppose the conditions of Theorem 4.10 are satisfied with $\widetilde{\gamma}_G = (1 - \delta/3)(\gamma_G + \theta L'/(L' + \theta)) - L\rho_y/2 > 0$, and the step length rules (4.18) replaced with*

$$\tau < \min\left\{\frac{2\delta}{L' + \theta + 6L\rho_y}, \frac{\sqrt{(1-\kappa)\widetilde{\gamma}_{F^*}/\widetilde{\gamma}_G}}{R_K}\right\}, \quad \sigma = \frac{\widetilde{\gamma}_G}{\widetilde{\gamma}_{F^*}}\tau, \quad \text{and} \quad \omega = \frac{1}{1 + 2\widetilde{\gamma}_G\tau}.$$

If $x \mapsto \nabla K(x)^* \widehat{y}$ is hypomonotone in X with the parameter $\theta > -L'$, then $\|u^i - \widehat{u}\|^2$ converges to zero with the rate $O(1/(1 + 2\widetilde{\gamma}_G\tau)^N)$.

Example 4.1 (Forward–backward splitting). Take $Y = \mathbb{R}$, $F(z) = z$, and $K \in C^1(X;\mathbb{R})$ convex with Lipschitz gradient. Then $F^* = \delta_{\{1\}}$, in particular $\widehat{y} = 1$, so the hypomonotonicity follows from the convexity of K. Since σ and ω have no effect in Algorithm 1.1, it reduces to conventional forward–backward splitting, consisting of the single update $x^{i+1} := (I + \tau\partial G)^{-1}(x^i - \tau\nabla K(x^i))$.

In Lemma 4.20, we can take $\delta = 1$ and $\rho_y = 0$. Since now H is maximal monotone, we can apply Corollary 4.22 to obtain weak convergence under the standard condition $\tau L < 2$; see also [14]. Our other results can be used for linear and $O(1/N^2)$ convergence.

5 NUMERICAL EXAMPLES

We now illustrate the effects of acceleration together with the possibility of satisfying the assumptions on the step sizes using examples from [12]. As a nonlinear operator, we consider the mapping from a potential coefficient in an elliptic equation to the corresponding solution, i.e., for a Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$ and $X = Y = L^2(\Omega)$, we set $S : x \mapsto z$ for z satisfying

$$\begin{cases} \Delta z + xz = f & \text{on } \Omega, \\ \partial_\nu z = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.1}$$

Here $f \in L^2(\Omega)$ is given; for our examples below we take $f \equiv 1$. The operator S is uniformly bounded for all $x \geq \varepsilon > 0$ almost everywhere as well as completely continuous and twice Fréchet differentiable with uniformly bounded derivatives. Furthermore, for any $h \in X$, the application $\nabla S(x)^* h$ of the adjoint Fréchet derivative can be computed by solving a similar elliptic equation; see [12, Section 3]. For our numerical examples, we take $\Omega = (-1, 1)$ and approximate S by a standard finite element discretization on a uniform mesh with 1000 elements with piecewise constant x and piecewise linear z. We use the MATLAB codes accompanying [12] that can be downloaded from [11].

The first example is the L^1 fitting problem

$$\min_{x \in L^2(\Omega)} \frac{1}{\alpha}\|S(x) - z^\delta\|_{L^1} + \frac{1}{2}\|x\|_{L^2}^2, \tag{5.2}$$

for some noisy data $z^\delta \in L^2(\Omega)$ and a regularization parameter $\alpha > 0$; see [12, Section 3.1] for details. For the purpose of this example, we take z^δ as arising from random-valued impulsive noise applied to $z^\dagger = S(x^\dagger)$ for $x^\dagger(t) = 2 - |t|$ and $\alpha = 10^{-2}$. This fits into the framework of problem (P) with $F(y) = \frac{1}{\alpha}\|y\|_{L^1}$, $G(x) = \frac{1}{2}\|x\|_{L^2}^2$, and $K(x) = S(x) - z^\delta$. (Note that in contrast to [12], we do not introduce a Moreau–Yosida regularization of F here.) Due to the properties of S, Assumption 4.1 are satisfied with $\theta \geq -L$, $\gamma_G = 1$ and $\gamma_F = 0$. As in [12], we estimate the Lipschitz constant L by $\widetilde{L} = \max\{1, \|\nabla S'(u^0)u^0\|/\|u^0\|\} \approx 1$. We then set $\tau_0 = (4\widetilde{L})^{-1}$ and $\sigma_0 = (2\widetilde{L})^{-1}$. The starting points are chosen as $x_0 \equiv 1$ and $y_0 \equiv 0$. Figure 1 shows the convergence behavior $\|x^N - \hat{x}\|_{L^2}^2$ of the primal iterates for $N \in \{1, \ldots, N_{\max}\}$ for $N_{\max} = 10^4$, both without and with acceleration. Since the exact minimiser to (5.2) is unavailable, here we take $\hat{x} := x^{2N_{\max}}$ as an approximation. As can be seen, the convergence in the first case (corresponding to $\widetilde{\gamma}_G = 0$) is

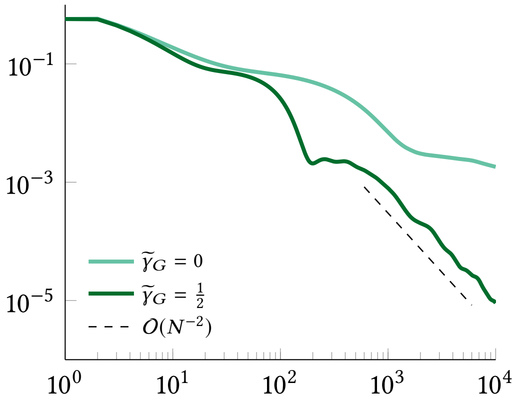

Figure 1: L^1 fitting: $\|x^N - \hat{x}\|_{L^2}^2$ for different values of $\widetilde{\gamma}_G$

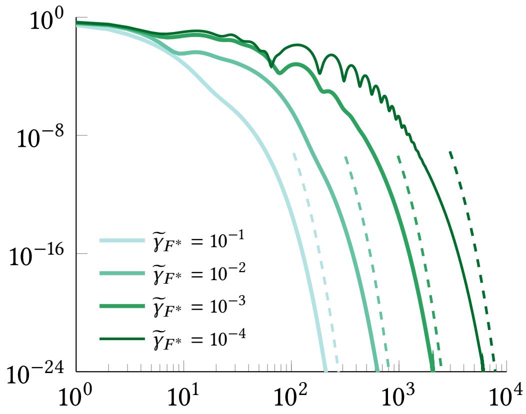

Figure 2: L^1 fitting: $\|u^N - \hat{u}\|_{L^2\times L^2}^2$ (solid) and bounds $(1+2\widetilde{\gamma}_G\tau)^{-N}$ (dashed) for strongly convex F^* and different values of $\widetilde{\gamma}_{F^*}$

at best $O(1/N)$, while the accelerated algorithm according to Theorem 4.7 with $\widetilde{\gamma}_G = \frac{1}{2} < \gamma_G$ indeed eventually enters a region with quadratic convergence. If we replace F by its Moreau–Yosida regularization F_γ, i.e., replace F^* by $F_\gamma^* := F^* + \frac{\gamma}{2}\|\cdot\|_Y^2$, Theorem 4.10 is applicable for $\widetilde{\gamma}_{F^*} = \gamma > 0$. As Figure 2 shows for different choices of γ and constant step sizes $\tau = \sqrt{\widetilde{\gamma}_{F^*}/\widetilde{\gamma}_G}\widetilde{L}^{-1}$, $\sigma = (\widetilde{\gamma}_G/\widetilde{\gamma}_{F^*})\tau$, the corresponding algorithm leads to linear convergence of the full iterates $\|u^N - \hat{u}\|_{L^2\times L^2}^2$ with a rate of $(1+2\widetilde{\gamma}_G\tau)^{-N}$ (which depends on γ by way of τ).

We also consider the example of optimal control with state constraints mentioned in the introduction, i.e.,

$$\min_{x\in L^2} \frac{1}{2\alpha}\|S(x) - z^d\|_{L^2}^2 + \frac{1}{2}\|x\|_{L^2}^2 \quad \text{s.t.} \quad [S(x)](t) \le c \quad \text{a.e. in } \Omega, \tag{5.3}$$

see [12, Section 3.3] for details. Here we choose $z^d = S(x^\dagger)$ with $x^\dagger$ as above, $\alpha = 10^{-3}$, and $c = 0.68$ such that the state constraints are violated for z^d. Again, this fits into the framework of problem (P) with $F(y) = \frac{1}{2\alpha}\|y - z^d\|_{L^2}^2 + \delta_{(-\infty,c]}(y)$, $G(x) = \frac{1}{2}\|x\|_{L^2}^2$, and $K(x) = S(x)$. With the same parameter choice as in the last example, we again observe locally quadratic convergence for the accelerated algorithm (see Figure 3) as well as linear convergence if the state constraints are replaced by a Moreau–Yosida regularization (see Figure 4).

6 CONCLUSIONS

We have applied the testing framework, gradually developed in [22, 24, 25], to obtain sufficient conditions on primal and dual step lengths that ensure convergence and fast convergence rates of the NL-PDHGM. We have shown how usual acceleration rules give local $O(1/N^2)$ convergence, justifying their use in previously published numerical examples [12]. Moreover, we have provided novel linear convergence results, and demonstrated their usefulness in practice. These results are based on bounds on initial step lengths. We have further demonstrated how hypomonotonicity of the saddle term gradient can be used to obtain weaker bounds, indeed deriving standard

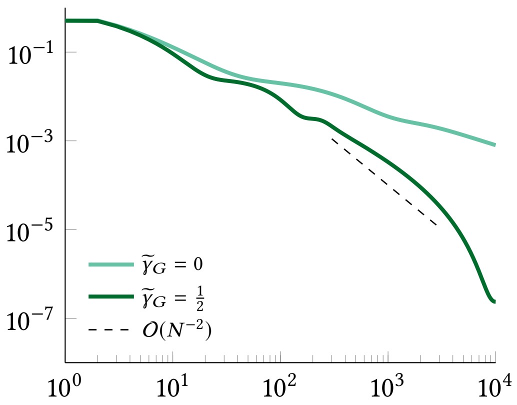

Figure 3: State constraints: $\|x^N - \hat{x}\|_{L^2}^2$ for different values of $\widetilde{\gamma}_G$

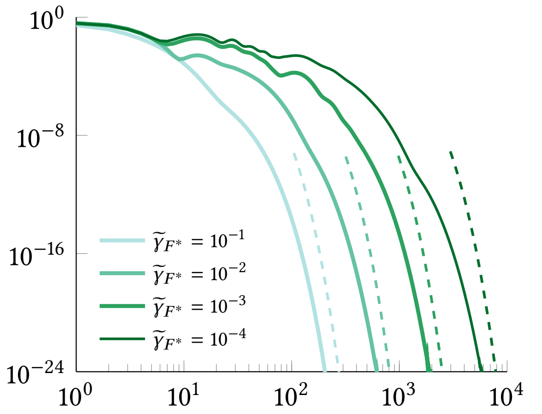

Figure 4: State constraints: $\|u^N - \hat{u}\|_{L^2\times L^2}^2$ (solid) and bounds $(1 + 2\widetilde{\gamma}_G\tau)^{-N}$ (dashed) for strongly convex F^* and different values of $\widetilde{\gamma}_{F^*}$

results for forward–backward splitting via this route. Since our main derivations were for general operators, a potential extension of the present work is to combine with [22] to derive block-coordinate methods for nonconvex problems.

ACKNOWLEDGEMENTS

T. Valkonen and S. Mazurenko have been supported by the EPSRC First Grant EP/P021298/1, “PARTIAL Analysis of Relations in Tasks of Inversion for Algorithmic Leverage”. C. Clason is supported by the German Science Foundation (DFG) under grant Cl 487/1-1.

A DATA STATEMENT FOR THE EPSRC

All data and source codes will be publicly deposited when the final accepted version of the manuscript is submitted.

APPENDIX A A SMALL IMPROVEMENT OF OPIAL’S LEMMA

The earliest version of the next lemma is contained in the proof of [19, Theorem 1].

Lemma A.1 ([6, Lemma 6]). *On a Hilbert space X, let $\hat{X} \subset X$ be closed and convex, and $\{x^i\}_{i\in\mathbb{N}} \subset X$. Then $x^i \rightharpoonup x^*$ weakly in X for some $x^* \in \hat{X}$ if:*

(i) $i \mapsto \|x^i - x^*\|$ is nonincreasing for all $x^* \in \hat{X}$.

(ii) All weak limit points of $\{x^i\}_{i\in\mathbb{N}}$ belong to $\hat{X}$.

We can improve it to the following:

Lemma A.2. *Let X be a Hilbert space, $\hat{X} \subset X$ (not necessarily closed or convex), and $\{x^i\}_{i\in\mathbb{N}} \subset X$. Also let $A_i \in \mathcal{L}(X;X)$ be self-adjoint and $A_i \geq \hat{\epsilon}^2 I$ for some $\hat{\epsilon} \neq 0$ for all $i \in \mathbb{N}$. If the following conditions hold, then $x^i \rightharpoonup x^*$ weakly in X for some $x^* \in \hat{X}$:*

(i) $i \mapsto \|x^i - \hat{x}\|_{A_i}$ *is nonincreasing for* some $\hat{x} \in \hat{X}$.

(ii) All weak limit points of $\{x^i\}_{i\in\mathbb{N}}$ belong to $\hat{X}$.

(iii) There exists C such that $\|A_i\| \leq C^2$ for all i, and for any weakly convergent subsequence x_{i_k} there exists $A_\infty \in \mathcal{L}(X;X)$ such that $A_{i_k}x \to A_\infty x$ strongly in X for all $x \in X$.

Proof. For $x \in \operatorname{cl}\operatorname{conv}\hat{X}$, define $p(x) := \liminf_{i\to\infty} \|x - x^i\|_{A_i}$. Clearly (i) yields

$$p(\hat{x}) = \lim_{i\to\infty} \|\hat{x} - x^i\|_{A_i} \in [0, \infty).$$

Using the triangle inequality and (iii), for any $x, x' \in \operatorname{cl}\operatorname{conv}\hat{X}$ moreover

$$0 \leq p(x) \leq p(x') + \limsup_{i\to\infty} \|x' - x\|_{A_i} \leq p(x') + C\|x' - x\|. \tag{A.1}$$

Choosing $x' = \hat{x}$ we see from (A.1) that p is well-defined and finite. It is moreover bounded from below. Given $\epsilon > 0$, we can therefore find $x^*_\epsilon \in \operatorname{cl}\operatorname{conv}\hat{X}$ such that $p(x^*_\epsilon)^2 - \epsilon^2 \leq \inf_{\operatorname{cl}\operatorname{conv}\hat{X}} p^2$. The norm $\|x^*_\epsilon\|$ is bounded from above for small values of ϵ: for the subsequence $\{x_{i_k}\}$ realizing the limes inferior in $p(x^*_\epsilon)$,

$$\|x^*_\epsilon\|_{A_{i_k}} \leq \|x^*_\epsilon - x^{i_k}\|_{A_{i_k}} + \|x^{i_k} - \hat{x}\|_{A_{i_k}} + \|\hat{x}\|_{A_{i_k}},$$

and consequently

$$\hat{\epsilon}\|x^*_\epsilon\| \leq \left(\inf_{\operatorname{cl}\operatorname{conv}\hat{X}} p\right) + \epsilon + \|x^0 - \hat{x}\|_{A_0} + C\|\hat{x}\|,$$

so there is a subsequence of $\|x^*_\epsilon\|$ weakly converging to some x^* when $\epsilon \searrow 0$. Without loss of generality, by restricting the allowed values of ϵ, we may assume that x^* is unique.

Let x^{**} be some weak limit of $\{x^i\}$. By (ii), $x^{**} \in \hat{X}$. We have to show that $x^* = x^{**}$. For simplicity of notation, we may assume that the whole sequence $\{x^i\}$ converges weakly to x^{**}. By (iii), for any $x \in X$, we have

$$\lim_{i\to\infty} \langle x, x^*_\epsilon - x^i\rangle_{A_i} = \lim_{i\to\infty} \left(\langle x, x^*_\epsilon - x^i\rangle_{A_\infty} + \langle (A_i - A_\infty)x, x^*_\epsilon - x^i\rangle\right) = \langle x, x^*_\epsilon - x^{**}\rangle_{A_\infty}. \tag{A.2}$$

Moreover, for any $\lambda \in (0, 1)$, we have $x^*_{\epsilon,\lambda} := (1 - \lambda)x^*_\epsilon + \lambda x^{**} \in \operatorname{cl}\operatorname{conv}\hat{X}$. Now, since x^* is a minimizer of p on $\operatorname{cl}\operatorname{conv}\hat{X}$, we estimate

$$\begin{aligned} p(x^*_\epsilon)^2 - \epsilon^2 \leq p(x^*_{\epsilon,\lambda})^2 &= p(x^*_\epsilon)^2 + \lim_{i\to\infty}\left(\lambda^2\|x^*_\epsilon - x^{**}\|^2_{A_i} - 2\lambda\langle x^*_\epsilon - x^{**}, x^*_\epsilon - x^i\rangle_{A_i}\right) \\ &= p(x^*_\epsilon)^2 + (\lambda^2 - 2\lambda)\|x^*_\epsilon - x^{**}\|^2_{A_\infty}. \end{aligned} \tag{A.3}$$

In the second equality we have used (iii) and (A.2). Now, since $\lambda^2 \leq 2\lambda$, we obtain

$$0 \leq (2\lambda - \lambda^2)\|x^*_\epsilon - x^{**}\|^2_{A_\infty} \leq \epsilon^2.$$

This implies $x^*_\epsilon \to x^{**}$ strongly as $\epsilon \searrow 0$. But also $x^*{}_\epsilon \rightharpoonup x^*$. Therefore $x^{**} = x^*$.

Finally, by $A_i \geq \hat{\epsilon} I$ and (i), the sequence $\{x^i\}$ is bounded, so any subsequence contains a weakly convergent subsequence. Since the limit is always x^*, the whole sequence converges weakly to x^*. □

Remark A.3. *The condition $A_i \geq \hat{\epsilon} I$ is implied if we replace (iii) by $A_i \to A_\infty$ in the operator topology with $A_\infty \geq 2\hat{\epsilon} I$.*

REFERENCES

[1] Aragón Artacho & Geoffroy, Metric subregularity of the convex subdifferential in Banach spaces, *J. Nonlinear Convex Anal.* 15 (2014), 35–47, arxiv: 1303.3654.

[2] Bachmayr & Burger, Iterative total variation schemes for nonlinear inverse problems, *Inverse Problems* 25 (2009), doi: 10.1088/0266-5611/25/10/105004.

[3] Bauschke & Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2017, doi: 10.1007/978-3-319-48311-5.

[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*, Springer International Publishing, 2016, 117–126, doi: 10.1007/978-3-319-55795-3_10, arxiv: 1511.00425,

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

[6] Browder, Nonexpansive nonlinear operators in a Banach space, *Proceedings of the National Academy of Sciences of the United States of America* 54 (1965), 1041, url: www.jstor.org/stable/73047.

[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 & Lions, Image recovery via total variation minimization and related problems, *Numerische Mathematik* 76 (1997), 167–188, doi: 10.1007/s002110050258.

[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] Cibulka, Dontchev & Kruger, Strong metric subregularity of mappings in variational analysis and optimization, *Journal of Mathematical Analysis and Applications* 457 (2018), 1247–1282, doi: j.jmaa.2016.11.045.

[11] Clason & Valkonen, Codes supporting: Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization [nlpdegm], 2017, doi: 10.5281/zenodo.398822.

[12] Clason & Valkonen, Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization, *SIAM Journal on Optimization* 27 (2017), 1313–1339, doi: 10.1137/16M1080859, arxiv: 1606.06219,

[13] Clason & Valkonen, Stability of saddle points via explicit coderivatives of pointwise subdifferentials, *Set-valued and Variational Analysis* 25 (2017), 69–112, doi: 10.1007/s11228-016-0366-7, arxiv: 1509.06582,

[14] Combettes, Solving monotone inclusions via compositions of nonexpansive averaged operators, *Optimization* 53 (2004), 475–504, doi: 10.1080/02331930412331327157.

[15] Dontchev & Rockafellar, *Implicit Functions and Solution Mappings: A View from Variational Analysis*, Springer New York, 2014, doi: 10.1007/978-1-4939-1037-3.

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

[17] Gfrerer, On directional metric regularity, subregularity and optimality conditions for nonsmooth mathematical programs, *Set-valued and Variational Analysis* 21 (2013), 151–176, doi: 10.1007/s11228-012-0220-5.

[18] Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, Springer International Publishing, 2017, doi: 10.1007/978-3-319-64277-2.

[19] Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bulletin of the American Mathematical Society* 73 (1967), 591–597, doi: 10.1090/S0002-9904-1967-11761-0.

[20] Rockafellar & Wets, *Variational Analysis*, Springer, 1998, doi: 10.1007/978-3-642-02431-3.

[21] Valkonen, A primal-dual hybrid gradient method for non-linear operators with applications to MRI, *Inverse Problems* 30 (2014), 055012, doi: 10.1088/0266-5611/30/5/055012, arxiv: 1309.5032,

[22] Valkonen, Block-proximal methods with spatially adapted acceleration, *Optimization* (2017), arxiv: 1609.07373, url: tuomov.iki.fi/m/blockcp.pdf.

[23] Valkonen, Preconditioned proximal point methods and notions of partial subregularity, 2017, arxiv: 1711.05123, url: tuomov.iki.fi/m/subreg.pdf.

[24] Valkonen, Testing and non-linear preconditioning of the proximal point method, 2017, arxiv: 1703.05705, url: tuomov.iki.fi/m/proxtest.pdf.

[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] Zhu & Valkonen, Spherical function regularization for parallel MRI reconstruction, 2018, arxiv: 1802.01108, url: tuomov.iki.fi/m/spherical_parmri.pdf.