arXiv:2211.04807v2 [math.OC] 20 Oct 2023

A NONSMOOTH PRIMAL-DUAL METHOD WITH INTERWOVEN PDE CONSTRAINT SOLVER

Bjørn Jensen[*] Tuomo Valkonen[†]

Abstract We introduce an efficient first-order primal-dual method for the solution of nonsmooth PDE-constrained optimization problems. We achieve this efficiency through *not* solving the PDE or its linearisation on each iteration of the optimization method. Instead, we run the method interwoven with a simple conventional linear system solver (Jacobi, Gauss–Seidel, conjugate gradients), always taking only *one step* of the linear system solver for each step of the optimization method. The control parameter is updated on each iteration as determined by the optimization method. We prove linear convergence under a second-order growth condition, and numerically demonstrate the performance on a variety of PDEs related to inverse problems involving boundary measurements.

1 INTRODUCTION

Our objective is to develop efficient first-order algorithms for the solution of PDE-constrained optimization problems of the type

$$\min_{x,u} F(x) + Q(u) + G(Kx) \quad \text{subject to} \quad B(u, w; x) = Lw \quad \text{for all} \quad w,$$

where K is a linear operator and the functions F, G, and Q are convex but the first two possibly nonsmooth. The functionals B and L model a partial differential equation in weak form, parametrised by x; for example, $B(u, w; x) = \langle \nabla u, x\nabla w\rangle$.

Semismooth Newton methods [25, 27] are conventionally used for such problems when a suitable reformulation exists [17, 19, 31, 32, 18]. Reformulations may not always be available, or yield effective algorithms. The solution of large linear systems may also pose scaling challenges. Therefore, first-order methods for PDE-constrained optimization have been proposed [6, 4, 24, 5] based on the primal-dual proximal splitting (PDPS) of [3]. The original version applies to convex problems of the form

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

The primal-dual expansion permits efficient treatment of $G \circ K$ for nonsmooth G. In [6, 4, 24, 5] K may be nonlinear, such as the solution operator of a nonlinear PDE.

However, first-order methods generally require a very large number of iterations to exhibit convergence. If the iterations are cheap, they can, nevertheless, achieve good performance. If the iterations are expensive, such as when a PDE needs to be solved on each step, their performance can be poor. Therefore, especially in inverse problems research, Gauss–Newton -type approaches are common for (1.1)

This research was supported by the Academy of Finland decisions 314701, 320022, and 345389.

[*] Department of Mathematical Information Technology, University of Jyväskylä, Finland bjorn.c.s.jensen@jyu.fi

[†] ModeMat, Escuela Politécnica Nacional, Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland tuomo.valkonen@iki.fi

with nonlinear K; see, e.g., [8, 35, 20]. They are easy: first linearise K, then apply a convex optimization method or, in simplest cases, a linear system solver. Repeat. Even when a first-order method is used for the subproblem, Gauss–Newton methods can be significantly faster than full first-order methods [20] if they converge at all [33]. This stems from the following and only practical difference between the PDPS for nonlinear K and Gauss–Newton applied to (1.1) with PDPS for the inner problems: the former re-linearizes and factors K on *each* PDPS iteration, the latter only on each outer Gauss–Newton iteration.

In this work, we avoid forming and factorizing the PDE solution operators altogether by *running an iterative solver for the constantly adapting PDE simultaneously with the optimization method.* This may be compared to the approach to bilevel optimization in [29]. We concentrate on the simple Jacobi and Gauss–Seidel splitting methods for the PDE, while the optimization method is based on the PDPS, as we describe in Section 2. We prove convergence in Section 3 using the testing approach introduced in [34] and further elucidated in [7]. We explain how standard splittings and PDEs fit into the framework in Section 4, and finish with numerical experiments in Section 5.

Pseudo-time-stepping one-shot methods have been introduced in [30] and further studied, among others, in [28, 21, 14, 13, 12, 1, 11, 15]. A "one-shot" approach, as opposed to an "all-at-once" approach, solves the PDE constraints on each step, instead of considering them part of a unified system of optimality conditions. The aforementioned works solve these constraints inexactly through "pseudo-"time-stepping. This corresponds to the trivial split $A_x = (A_x - \mathrm{Id}) + \mathrm{Id}$ where A_x is such that $\langle A_x u, w\rangle = B(u, w; x)$. We will, instead, apply Jacobi, Gauss–Seidel or even (quasi-)conjugate gradient splitting on A_x. In [11, 1] Jacobi and Gauss–Seidel updates are used for the control variable, but not for the PDEs. The authors of [15] come closest to introducing non-trivial splitting of the PDEs via Hessian approximation. However, they and the other aforementioned works generally restrict themselves to smooth problems and employ gradient descent, Newton-type methods, or sequential quadratic programming (SQP) for the control variable x. Our focus is on nonsmooth problems involving, in particular, total variation regularization $G(Kx) = \|\nabla x\|_1$.

NOTATION AND BASIC RESULTS

Let X be a normed space. We write $\langle \,\cdot\,|\,\cdot\,\rangle$ for the dual product and, in a Hilbert space, $\langle \,\cdot\,, \,\cdot\,\rangle$ for the inner product. The order of the arguments in the dual product is not important when the action is obvious from context. For X a Hilbert space, we denote by $\mathrm{In}_X : X \hookrightarrow X^*$ the canonical injection, $\langle \mathrm{In}_X\, x|\tilde{x}\rangle = \langle x, \tilde{x}\rangle$ for all $x, \tilde{x} \in X$.

We write $\mathbb{L}(X; Y)$ for the space of bounded linear operators between X and Y. We write $\mathrm{Id}_X = \mathrm{Id} \in \mathbb{L}(X; X)$ for the identity operator on X. If $M \in \mathbb{L}(X; X^*)$ is non-negative and self-adjoint, i.e., $\langle Mx|y\rangle = \langle x|My\rangle$ and $\langle x|Mx\rangle \geq 0$ for all $x, y \in X$, we define $\|x\|_M := \sqrt{\langle x|Mx\rangle}$. Then the *three-point identity* holds:

$$\langle M(x - y)|x - z\rangle = \frac{1}{2}\|x - y\|_M^2 - \frac{1}{2}\|y - z\|_M^2 + \frac{1}{2}\|x - z\|_M^2 \qquad \text{for all } x, y, z \in X. \tag{1.2}$$

We extensively use the vector Young's inequality

$$\langle x|y\rangle \leq \frac{1}{2a}\|x\|_X^2 + \frac{a}{2}\|y\|_{X^*}^2 \quad (x \in X,\ y \in X^*,\ a > 0). \tag{1.3}$$

These expressions hold in Hilbert spaces also with the inner product in place of the dual product. We write $M^\star$ for the inner product adjoint of M, and M^* for the dual product adjoint.

We write $\operatorname{dom} F$ for the effective domain, and F^* for the Fenchel conjugate of $F : X \to \overline{\mathbb{R}} := [-\infty, \infty]$. We write $F'(x) \in X^*$ for the Fréchet derivative at x when it exists, and, if X is a Hilbert space, $\nabla F(x) \in X$ for its Riesz presentation. For convex F on a Hilbert space X, we write $\partial F(x) \subset X$ for the subdifferential at $x \in X$ (or, more precisely, the corresponding set of Riesz representations, but aside from a single proof in Appendix A, we will not be needing subderivatives as elements of X^*). We then define the proximal map

$$\operatorname{prox}_F(x) := (\operatorname{Id} + \partial F)^{-1}(x) = \operatorname*{arg\,min}_{\tilde{x} \in X} \left\{ F(\tilde{x}) + \frac{1}{2}\|\tilde{x} - x\|_X^2 \right\}, \quad x \in X.$$

We denote the $\{0, \infty\}$-valued indicator function of a set A by δ_A.

We occasionally apply operations on $x \in X$ to all elements of sets $A \subset X$, writing $\langle x + A|z\rangle := \{\langle x + a|z\rangle \mid a \in A\}$. For $B \subset \mathbb{R}$, we write $B \geq c$ if $b \geq c$ for all $b \in B$.

On a Lipschitz domain $\Omega \subset \mathbb{R}^n$, we write $\operatorname{trace}_{\partial\Omega} \in \mathbb{L}(H^1(\Omega); L^2(\partial\Omega))$ for the trace operator on the boundary $\partial\Omega$.

2 PROBLEM AND PROPOSED ALGORITHM

We start by introducing in detail the type of problem we are trying to solve. We then rewrite in Section 2.1 its optimality conditions in a form suitable for developing our proposed method in Section 2.3. Before this we recall the structure and derivation of the basic PDPS in Section 2.2.

2.1 PROBLEM DESCRIPTION

Our objective is to solve

$$\min_x J(x) := F(x) + Q(S(x)) + G(Kx), \tag{2.1}$$

where $F : X \to \overline{\mathbb{R}}$, $G : Y \to \overline{\mathbb{R}}$, and $Q : U \to \mathbb{R}$ are convex, proper, and lower semicontinuous on Hilbert spaces X, U, and Y with Q Fréchet differentiable. We assume $K \in \mathbb{L}(X; Y)$ while $S : X \ni x \mapsto u \in U$ is a solution operator of the weak PDE

$$B(u, w; x) = Lw \quad \text{for all} \quad w \in W. \tag{2.2}$$

Here $L \in U^*$ and $B : U \times W \times X \to \mathbb{R}$ is continuous, and affine-linear-affine in its three arguments. The space W is Hilbert, possibly distinct from U to model nonhomogeneous boundary conditions. For this initial development, we will tacitly assume unique $S(x)$ and $\nabla S(x)$ to exist for all $x \in \operatorname{dom} F$, but later on in the manuscript, do not directly impose this restriction, or use S.

Example 2.1 (A linear PDE). On a Lipschitz domain $\Omega \subset \mathbb{R}^n$, consider the PDE

$$\begin{cases} \nabla \cdot \nabla u = x, & \text{on } \Omega, \\ u = g, & \text{on } \partial\Omega. \end{cases}$$

For the weak form (2.2) we can take the spaces $U = H^1(\Omega)$, $W = H_0^1(\Omega) \times H^{1/2}(\partial\Omega)$, and $X = L^2(\Omega)$. Writing $w = (w_\Omega, w_\partial)$, we then set

$$B(u, w; x) = \langle \nabla u, \nabla w_\Omega \rangle_{L^2(\Omega)} - \langle x, w_\Omega \rangle_{L^2(\Omega)} + \langle \operatorname{trace}_{\partial\Omega} u, w_\partial \rangle_{L^2(\partial\Omega)} \quad \text{and} \quad Lw := \langle g, w_\partial \rangle_{L^2(\partial\Omega)}.$$

Example 2.2 (A nonlinear PDE). On a Lipschitz domain $\Omega \subset \mathbb{R}^n$, consider the PDE

$$\begin{cases} \nabla \cdot (x \nabla u) = 0, & \text{on } \Omega, \\ u = g, & \text{on } \partial\Omega. \end{cases}$$

For the weak form (2.2) we can take the spaces $U \subset H^1(\Omega)$, $W \subset H_0^1(\Omega) \times H^{1/2}(\partial\Omega)$, and $X \subset L^2(\Omega)$, such that at least one of these subspaces ensures the corresponding x, ∇u, or ∇w to be in the relevant

L^∞ space. This, in practise, requires one of the subspaces to be finite-dimensional, or X to be $H^k(\Omega)$ for $k > n/2$, such that the boundedness of Ω and Sobolev's inequalities provide the L^∞ bound. The latter is an option in infinite-dimensional theory, but in finite-dimensional realisations, it is desirable to use a standard 2-norm in X, as proximal operators and gradient steps with respect to H^k-norms (for $k > 0$) are computationally expensive. Writing $w = (w_\Omega, w_\partial)$, we then set

$$B(u, w; x) = \langle x\nabla u, \nabla w_\Omega\rangle_{L^2(\Omega)} + \langle \operatorname{trace}_{\partial\Omega} u, w_\partial\rangle_{L^2(\partial\Omega)} \quad \text{and} \quad Lw := \langle g, w_\partial\rangle_{L^2(\partial\Omega)}.$$

To ensure the coercivity of $B(\,\boldsymbol{\cdot}\,, \,\boldsymbol{\cdot}\,; x)$, and hence the existence of unique solutions to (2.2), we will further need to restrict x through $\operatorname{dom} F$.

We require the sum and chain rules for convex subdifferentials to hold on $F + G \circ K$. This is the case when

$$\text{there exists an } x \in \operatorname{dom}(G \circ K) \cap \operatorname{dom} F \text{ with } Kx \in \operatorname{int}(\operatorname{dom} G). \tag{2.3}$$

We refer to [7] for basic results and concepts of infinite-dimensional convex analysis. Then by the Fréchet differentiability of Q and the compatibility of limiting (Mordukhovich) subdifferentials (denoted ∂_M) with Fréchet derivatives and convex subdifferentials [26, 7],

$$\partial_M J(x) = \partial F(x) + \nabla S(x)^\star \nabla Q(S(x)) + K^\star \partial G(Kx).$$

Therefore, the Fermat principle for limiting subdifferentials and simple rearrangements (see [33, 4] or [7, Chapter 15]) establish for (2.1) in terms of $(\bar u, \bar w, \bar x, \bar y) \in U \times W \times X \times Y$ the necessary first-order optimality condition

$$\begin{cases} \bar u = S(\bar x), \\ -\nabla S(\bar x)^\star \nabla Q(\bar u) - K^\star \bar y \in \partial F(\bar x), \\ K\bar x \in \partial G^*(\bar y). \end{cases} \tag{2.4}$$

We recall that $G^* : Y \to \overline{\mathbb{R}}$ is the Fenchel conjugate of G.

The term $\nabla S(\bar x)^\star \nabla Q(\bar u)$ involves the solution $\bar u$ to the original PDE and the solution $\bar w$ to an adjoint PDE. We derive it from a primal-dual reformulation of (2.1). To do this, we first observe that since B is affine in x, it can be decomposed as

$$B(u, w; x) = B_x(u, w; x) + B_{\text{const}}(u, w), \tag{2.5}$$

where, $B_x : U \times W \times X \to \mathbb{R}$ is affine-linear-linear, and $B_{\text{const}} : U \times W \to \mathbb{R}$ is affine-linear. Indeed $B_{\text{const}}(u, w) = B(u, w; 0)$, and $B_x(u, w; x) = B(u, w; x) - B(u, w; 0)$. We then introduce the Riesz representation $\bar\nabla_x B(u, w)$ of $B_x(u, w; \,\boldsymbol{\cdot}\,) \in X^*$. Thus

$$\langle \bar\nabla_x B(u, w), x\rangle_X = B_x(u, w; x) \quad \text{for all } u \in U,\ w \in W,\ x \in X. \tag{2.6}$$

We have $\nabla_x B(u, w; x) \equiv \bar\nabla_x B(u, w) \in X$ for all $x \in X$.

Clearly, also, B_x is an abbreviation for $(u, w; x) \to D_x B(u, w, 0)(x)$, where, just here, we write D_x for the Fréchet derivative with respect to x. Likewise we write B_u to abbreviate $(u, w; x) \to D_u B(0, w, x)(u)$, and B_{xu} to abbreviate $(u, w; x) \to D_u B_x(0, w, x)(u)$. If B is linear in u, then $B_u = B$; and if B is linear in both u and x, then $B_{xu} = B$.

We may now write (2.1) as[1]

$$\min_{x,u} \max_{w}\ F(x) + Q(u) + B(u, w; x) - Lw + G(Kx) \tag{2.7}$$

[1] If the PDE (2.2) does not have a solution u for any $x \in \operatorname{dom} F \cap \operatorname{dom}(G \circ K)$, the inner "max" will be infinite, not reached, and technically, therefore, a "sup". In this case also (2.1) has no solution. If (2.1) has a solution, there must exist some (x, u) for which (any) w reaches the "max". Likewise, y reaching the corresponding "max" exists for any $x \in \operatorname{dom}(G \circ K)$ by basic properties of Fenchel conjugates of convex, proper, lower semicontinuous functions.

or

$$\min_{x,u}\max_{w,y}\; F(x) + Q(u) + B(u,w;x) - Lw + \langle Kx, y\rangle_Y - G^*(y). \tag{2.8}$$

In terms of $(\bar u, \bar w, \bar x, \bar y) \in U \times W \times X \times Y$, subject to a qualification condition, this problem has the necessary first-order optimality conditions

$$\begin{cases} B(\bar u, \tilde w; \bar x) = L\tilde w & \text{for all} \quad \tilde w \in W, \\ B_u(\tilde u, \bar w; \bar x) = -Q'(\bar u)\tilde u & \text{for all} \quad \tilde u \in U, \\ -\bar\nabla_x B(\bar u, \bar w) - K^\star \bar y \in \partial F(\bar x), & \\ K\bar x \in \partial G^*(\bar y). & \end{cases} \tag{2.9}$$

This is our principal form of optimality conditions for (2.1).

It is easy to see that (2.9) are necessary for $(\bar u, \bar w, \bar x, \bar y)$ to be a saddle point of (2.8). The next theorem shows, subject to qualification conditions, that (2.9) are also necessary for a solution to (2.8) (which may not be a saddle point in the non-convex-concave setting). Note that $w \in W$ is inconsequential in (2.8). If one choice forms a part of a solution of the problem, so does any other (or else the problem has no solution at all). However, $\bar w$ solving (2.9) is more precisely determined.

Theorem 2.3. *Suppose $(\bar u, w, \bar x, \bar y) \in U \times W \times X \times Y$ solve (2.8). If, moreover,* $\operatorname{int}\operatorname{dom}[F + G \circ K] \neq \emptyset$*, and, for some $c > 0$,*

$$\sup_{\|(h_x,h_u)\|=1} B_x(\bar u, w; h_x) + B_u(h_u, w; \bar x) \geq c\|w\| \quad \textit{for all} \quad w \in W \quad \textit{and} \tag{2.10a}$$

$$B_u(\tilde u, w; \bar x) = 0 \textit{ for all } \tilde u \implies B_x(\bar u, w; x) = 0 \textit{ for all } x \in \operatorname{dom}(F + G \circ K), \tag{2.10b}$$

then (2.9) holds for some $\bar w \in W$.

After an affine shift and restriction of x to a subspace, the condition $\operatorname{int}\operatorname{dom}[F+G\circ K] \neq \emptyset$ can always be relaxed to the corresponding relative interior being non-empty. Since the proof of Theorem 2.3 is long and depends on techniques not needed in our main line of work, we relegate it to Appendix A.

Example 2.4. If $W = U$, taking $h_u = w/\|w\|$ and $h_x = 0$, we see that the qualification conditions (2.10) hold when $B_u(\,\cdot\,, \,\cdot\,; \bar x)$ is coercive. Similarly, also when $W \neq U$, if the weak coercivity conditions of the Babuška–Lax–Milgram theorem hold for $(w, h_u) \mapsto B_u(h_u, w; \bar x)$, then so do (2.10).

The second line of (2.9) is the adjoint PDE, needed for $\nabla S(\bar x)^* \nabla Q(\bar u)$ in (2.4):

Corollary 2.5. *Suppose (2.10) hold for $\bar x = x \in X$, some $w \in W$, and $\bar u = u$ a unique solution to (2.2). Then the solution operator S of (2.2) satisfies for all $z \in U$ that*

$$\nabla S(x)^\star z = \bar\nabla_x B(u,w) \quad \textit{where} \quad u = S(x) \quad \textit{and} \quad \begin{cases} w \textit{ solves the weak adjoint PDE:} \\ B_u(\tilde u, w; x) = -\langle z, \tilde u\rangle \textit{ for all } \tilde u \in U. \end{cases}$$

Proof. Take $F \equiv 0$, $K = \mathrm{Id}$, $G \equiv \delta_{\{x\}}$, and $Q = \langle z, \,\cdot\,\rangle_U$. Then any solution $(\bar u, w, \bar x, y)$ to (2.8) has $\bar x = x$. Since $G^*(\tilde y) = \langle x, \tilde y\rangle$, any choice of y and w solve (2.8). Therefore, Theorem 2.3 applied to the problem we just constructed shows that

$$B_u(\tilde u, w; x) = -\langle z, \tilde u\rangle_U \text{ for all } \tilde u \in U \quad \text{and} \quad -\bar\nabla_x B(u,w) - y = 0.$$

On the other hand, (2.4) reduces to some y satisfying $-\nabla S(x)^\star z - y = 0$. Comparing these two expressions, we obtain the claim.

□

2.2 PRIMAL-DUAL PROXIMAL SPLITTING: A RECAP

The primal-dual proximal splitting (PDPS) for (1.1) is based on the optimality conditions

$$\begin{cases} -K^\star \bar{y} \in \partial F(\bar{x}), \\ \quad K\bar{x} \in \partial G^*(\bar{y}). \end{cases} \tag{2.11}$$

These are just the last two lines of (2.9) without $\bar{\nabla}_x B$. As derived in [34, 16, 7], the basic (unaccelerated) PDPS solves (2.11) by iteratively solving for each $k \in \mathbb{N}$ the system

$$\begin{cases} 0 \in \tau\partial F(x^{k+1}) + \tau K^\star y^k + x^{k+1} - x^k \\ 0 \in \sigma\partial G^*(y^{k+1}) - \sigma K[x^{k+1} + \omega(x^{k+1} - x^k)] + y^{k+1} - y^k, \end{cases} \tag{2.12}$$

where the primal and dual step length parameters $\tau, \sigma > 0$ satisfy $\tau\sigma\|K\| < 1$, and the over-relaxation parameter $\omega = 1$. We can write (2.12) in explicit form as

$$\begin{cases} x^{k+1} := \operatorname{prox}_{\tau F}(x^k - \tau K^\star y^k), \\ y^{k+1} := \operatorname{prox}_{\sigma G^*}\left(y^k + \sigma K[x^{k+1} + \omega(x^{k+1} - x^k)]\right). \end{cases}$$

2.3 ALGORITHM DERIVATION

The derivation of the PDPS and the optimality conditions (2.9) suggest to solve (2.9) by iteratively solving

$$\begin{cases} \quad B(u^{k+1}, \,\cdot\,; x^k) = L, \\ B_u(\,\cdot\,, w^{k+1}; x^k) = -Q'(u^{k+1}), \\ \qquad\qquad 0 \in \tau_k\partial F(x^{k+1}) + \tau_k\bar{\nabla}_x B(u^{k+1}, w^{k+1}) + \tau K^\star y^k + x^{k+1} - x^k \\ \qquad\qquad 0 \in \sigma_{k+1}\partial G^*(y^{k+1}) - \sigma_{k+1}K[x^{k+1} + \omega_k(x^{k+1} - x^k)] + y^{k+1} - y^k. \end{cases} \tag{2.13}$$

We have made the step length and over-relaxation parameters iteration-dependent for acceleration purposes. The indexing τ_k and σ_{k+1} is off-by-one to maintain the symmetric update rules from [3].

The method in (2.13) still requires exact solution of the PDEs. For some splitting operators $\Gamma_k, \Upsilon_k : U \times W \times X \to \mathbb{R}$, we therefore transform the first two lines into

$$B(u^{k+1}, \,\cdot\,; x^k) - \Gamma_k(u^{k+1} - u^k, \,\cdot\,; x^k) = L \quad \text{and} \tag{2.14a}$$

$$B_u(\,\cdot\,, w^{k+1}; x^k) - \Upsilon_k(\,\cdot\,, w^{k+1} - w^k; x^k) = -Q'(u^{k+1}). \tag{2.14b}$$

Example 2.6 (Splitting). Let $B(u, w; x) = \langle A_x u, w\rangle$ for symmetric $A_x \in \mathbb{R}^{n\times n}$ on $U = W = \mathbb{R}^n$. Take $\Gamma_k(u, w; x) = \langle [A_x - N_x]u, w\rangle$ and $\Upsilon_k = \Gamma_k$ for easily invertible $N_x \in \mathbb{R}^{n\times n}$. With $L = \langle b, \,\cdot\,\rangle$, $b \in \mathbb{R}^n$ and $M_x := A_x - N_x$, (2.14) now reads

$$N_{x^k}u^{k+1} = b - M_{x^k}u^k \quad \text{and} \quad N_{x^k}w^{k+1} = -\nabla Q(u^{k+1}) - M_{x^k}w^k. \tag{2.15}$$

For Jacobi splitting we take N_{x^k} as the diagonal part of A_{x^k}, and for Gauss–Seidel splitting as the lower triangle including the diagonal. We study these choices further in Section 4.2.

Let us introduce the general notation $v = (u, w, x, y)$ as well as the *step length operators* $T_k \in \mathbb{L}(U^* \times W^* \times X \times Y; U^* \times W^* \times X \times Y)$,

$$T_k := \operatorname{diag}\left(\mathrm{Id}_{U^*} \quad \mathrm{Id}_{W^*} \quad \tau_k\,\mathrm{Id}_X \quad \sigma_{k+1}\,\mathrm{Id}_Y\right), \tag{2.16}$$

Algorithm 2.1 Primal dual splitting with parallel adaptive PDE solves (PDPAP)

```
Require: $F : X \to \overline{\mathbb{R}}$, $G^* : Y \to \overline{\mathbb{R}}$, Fréchet-differentiable $Q : U \to \mathbb{R}$; $K \in \mathbb{L}(X; Y)$, $L \in U^*$; and
    $B : U \times W \times X \to \mathbb{R}$, bilinear in the first two variables, affine in the third, all on Hilbert spaces
    $X$, $Y$, $U$, and $W$. Riesz representation $\bar\nabla_x B(u, w)$ of $B_x(u, w; \,\cdot\,)$; see (2.6). For all $k \in \mathbb{N}$, splittings
    $\Gamma_k, \Upsilon_k : U \times W \times X \to \mathbb{R}$ and step length and over-relaxation parameters $\tau_k, \sigma_{k+1}, \omega_k > 0$; see
    Theorem 3.10 or 3.11.
Pick an initial iterate $(u^0, w^0, x^0, y^0) \in U \times W \times X \times Y$.
for $k \in \mathbb{N}$ do
    Solve $u^{k+1} \in U$ from the split weak PDE
        $B(u^{k+1}, \tilde w; x^k) - \Gamma_k(u^{k+1} - u^k, \tilde w; x^k) = L\tilde w \quad \text{for all} \quad \tilde w \in W.$
    Solve $w^{k+1} \in W$ from the split weak adjoint PDE
        $B_u(\tilde u, w^{k+1}; x^k) - \Upsilon_k(\tilde u, w^{k+1} - w^k; x^k) = -Q'(u^{k+1})\tilde u \quad \text{for all} \quad \tilde u \in U.$
    $x^{k+1} := \operatorname{prox}_{\tau_k F}\bigl(x^k - \tau_k \bar\nabla_x B(u^{k+1}, w^{k+1}) - \tau_k K^\star y^k\bigr)$
    $\bar x^{k+1} := x^{k+1} + \omega_k (x^{k+1} - x^k)$
    $y^{k+1} := \operatorname{prox}_{\sigma_{k+1} G^*}\bigl(y^k + \sigma_{k+1} K \bar x^{k+1}\bigr)$
end for
```

the set-valued operators $H_k : U \times W \times X \times Y \rightrightarrows U^* \times W^* \times X \times Y$,

$$H_k(v) := \begin{pmatrix} B(u, \,\cdot\,; x^k) - \Gamma_k(u - u^k, \,\cdot\,; x^k) - L \\ B_u(\,\cdot\,, w; x^k) - \Upsilon_k(\,\cdot\,, w - w^k; x^k) + Q'(u) \\ \partial F(x) + \bar\nabla_x B(u, w) + K^\star y \\ \partial G^*(y) - Kx \end{pmatrix}, \tag{2.17}$$

and the *preconditioning operators* $M_k \in \mathbb{L}(U \times W \times X \times Y; U^* \times W^* \times X \times Y)$,

$$M_k := \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \mathrm{Id}_X & -\tau_k K^\star \\ & & -\omega_k \sigma_{k+1} K & \mathrm{Id}_Y \end{pmatrix}. \tag{2.18}$$

The implicit form of our proposed algorithm for the solution of (2.1) is then

$$0 \in T_k H_k(v^{k+1}) + M_k(v^{k+1} - v^k). \tag{2.19}$$

Writing out (2.19) in terms of explicit proximal maps, we obtain Algorithm 2.1.

Remark 2.7. The index k for T_k, H_k, M_k in (2.16)–(2.19) is inconsistent with some of our earlier articles that would use the index $k+1$ similarly to the unknown v^{k+1}. We have decided to make this change to keep the notation lighter.

3 CONVERGENCE

We now treat the convergence of Algorithm 2.1. Following [34, 7] we "test" its implicit form (2.19) by applying on both sides the linear functional $\langle Z_k \,\cdot\, | v^{k+1} - \bar v\rangle$. Here Z_k is a convergence rate encoding "testing operator" (Section 3.2). A simple argument involving the three-point identity (1.2) and a growth estimate for H_k then yields in Section 3.3 a Féjer-type monotonicity estimate in terms of iteration-dependent norms. This establishes in Section 3.4 global convergence subject to a growth condition. We start with assumptions.

3.1 THE MAIN ASSUMPTIONS

We start with our main structural assumption. Further central conditions related to the PDE constraint will follow in Assumption 3.3, and through its verification for specific linear system solvers in Section 4.2.

Assumption 3.1 (Structure). On Hilbert spaces X, Y, U, and W, we are given convex, proper, and lower semicontinuous $F : X \to \overline{\mathbb{R}}$, $G^* : Y \to \overline{\mathbb{R}}$, and $Q : U \to \mathbb{R}$ with Q Fréchet differentiable, as well as $K \in \mathbb{L}(X; Y)$, $L \in U^*$, and $B : U \times W \times X \to \mathbb{R}$ affine-linear-affine. We assume:

(i) F and G are (strongly) convex with factors $\gamma_F, \gamma_{G^*} \geq 0$. With K they satisfy the condition (2.3) for the subdifferential sum and chain rules to be exact.

(ii) For all $x \in \operatorname{dom} F$, there exist solutions $(u, w) \in U \times W$ to the PDE $B(u, \boldsymbol{\cdot}; x) = L$ and the adjoint PDE $B_u(\boldsymbol{\cdot}, w; x) = -Q'(u)$.

We then fix a solution $\bar{v} = (\bar{u}, \bar{w}, \bar{x}, \bar{y}) \in U \times W \times X \times Y$ to (2.9) and assume that:

(iii) For some $\mathcal{S}(\bar{u}), \mathcal{S}(\bar{w}) \geq 0$, for all $(u, w) \in U \times W$ and $x \in \operatorname{dom} F$, we have

$$B_{xu}(u, \bar{w}; x - \bar{x}) \leq \sqrt{\mathcal{S}(\bar{w})}\|u\|_U\|x - \bar{x}\|_X \quad \text{and} \quad B_x(\bar{u}, w; x - \bar{x}) \leq \sqrt{\mathcal{S}(\bar{u})}\|w\|_W\|x - \bar{x}\|_X.$$

(iv) For some $C_x \geq 0$, for all $(u, w) \in U \times W$ and $x \in \operatorname{dom} F$ we have the bound

$$B_{xu}(u, w; x - \bar{x}) \leq C_x\|u\|_U\|w\|_W.$$

Remark 3.2. Part (i) is easy to check. In general, (iv) requires $\operatorname{dom} F$ to be bounded with respect to an ∞-norm with $B_x(u, w, x) \leq C\|u\|_U\|w\|_W\|x\|_\infty$ for some $C > 0$. Then $C_x = \sup_{x\in\operatorname{dom} F} C\|x\|_\infty$. If B_x is independent of u, i.e., for linear PDEs, both $C_x = 0$ and $\mathcal{S}(\bar{w}) = 0$, while $\mathcal{S}(\bar{u})$ is a constant independent of $\bar{u}$. We study (ii)–(iv) further in Section 4.1.

The next assumption encodes our conditions on the PDE splittings.

Assumption 3.3 (Splitting). Let Assumption 3.1 hold. For $k \in \mathbb{N}$, for which this assumption is to hold, we are given splitting operators $\Gamma_k, \Upsilon_k : U \times W \times X \to \mathbb{R}$ and $v^k = (u^k, w^k, x^k, y^k) \in U \times W \times X \times Y$ such that:

(i) Γ_k is linear in the second argument, Υ_k in the first.

(ii) There exist solutions u^{k+1} and w^{k+1} to the split equations (2.14).

(iii) For some $\gamma_B > 0$ and $C_Q, \pi_u, \pi_w \geq 0$, we have

$$\begin{aligned}\|u^k - \bar{u}\|_U^2 &\geq \gamma_B\|u^{k+1} - \bar{u}\|_U^2 - \pi_u\|x^k - \bar{x}\|_X^2 \quad \text{and}\\ \|w^k - \bar{w}\|_W^2 &\geq \gamma_B\|w^{k+1} - \bar{w}\|_W^2 - C_Q\|u^{k+1} - \bar{u}\|_U^2 - \pi_w\|x^k - \bar{x}\|_X^2.\end{aligned}$$

We verify the assumption for standard splittings in Section 4.2. The verification will introduce the assumption that Q' be Lipschitz. The Lipschitz factor then appears in C_Q, justifying the Q-subscript notation. Generally π_u and π_w model the x-sensitivity of B and B_u. For linear PDEs, such as Example 2.1, B_u does not depend on x. In that case most iterative solvers for the adjoint PDE would also be independent of x and have $\pi_w = 0$. The factor γ_B relates to the contractivity of the iterative solver.

The next, final, assumption introduces *testing parameters* that encode convergence rates and restrict the *step length parameters* in the standard primal-dual component of our method. It has no difference to the treatment of the PDPS in [34, 7]. Dependent on whether both, one, or none of $\tilde{\gamma}_F > 0$ and $\tilde{\gamma}_{G^*} > 0$, the parameters can be chosen to yield varying modes and rates of convergence.

Assumption 3.4 (Primal-dual parameters). Let Assumption 3.1 hold. For all $k \in \mathbb{N}$, the *testing parameters* $\varphi_k, \psi_k > 0$, *step length parameters* $\tau_k, \sigma_k > 0$, and the *over-relaxation parameter* $\omega_k \in (0,1]$ satisfy for some $\tilde\gamma_F \in [0, \gamma_F]$ and $\tilde\gamma_{G^*} \in [0, \gamma_{G^*}]$, and $\kappa \in (0,1)$ that

$$\begin{aligned} \varphi_{k+1} &= \varphi_k(1+2\tilde\gamma_F\tau_k), & \psi_{k+1} &= \psi_k(1+2\tilde\gamma_{G^*}\sigma_k), & & \\ \eta_k &:= \varphi_k\tau_k = \psi_k\sigma_k, & \omega_k &= \eta_{k+1}^{-1}\eta_k, \quad \text{and} & \kappa &\geq \frac{\tau_k\sigma_k}{1+2\tilde\gamma_{G^*}\sigma_k}\|K\|^2. \end{aligned}$$

3.2 THE TESTING OPERATOR

To complement the primal-dual testing parameters in Assumption 3.4, we introduce testing parameters $\lambda_k, \theta_k > 0$ corresponding to the PDE updates in our method; the first two lines of (2.19). We combine all of them into the *testing operator* $Z_k \in \mathbb{L}(U^* \times W^* \times X \times Y; U^* \times W^* \times X^* \times Y^*)$ defined by

$$Z_k := \operatorname{diag}\left(\lambda_k \operatorname{Id} \quad \theta_k \operatorname{Id} \quad \varphi_k \operatorname{In}_X \quad \psi_{k+1} \operatorname{In}_Y\right). \tag{3.1}$$

Recalling M_k and Z_k from (2.18) and (3.1), thanks to Assumption 3.4, we have

$$Z_kM_k = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \varphi_k \operatorname{In}_X & -\eta_k \operatorname{In}_X K^\star \\ & & -\eta_k \operatorname{In}_Y K & \psi_{k+1} \operatorname{In}_Y \end{pmatrix}. \tag{3.2}$$

Therefore,

$$Z_k(M_k + \Xi_k) = Z_{k+1}M_{k+1} + D_{k+1} \tag{3.3}$$

for skew-symmetric

$$D_{k+1} := \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 0 & (\eta_{k+1}+\eta_k) \operatorname{In}_X K^\star \\ & & -(\eta_{k+1}+\eta_k) \operatorname{In}_Y K & 0 \end{pmatrix}$$

and $\Xi_k \in \mathbb{L}(U \times W \times X \times Y; U^* \times W^* \times X^* \times Y^*)$ satisfying

$$Z_k\Xi_k = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 2\eta_k\tilde\gamma_F \operatorname{In}_X & 2\eta_k \operatorname{In}_X K^\star \\ & & -2\eta_{k+1} \operatorname{In}_Y K & 2\eta_{k+1}\tilde\gamma_{G^*} \operatorname{In}_Y \end{pmatrix}. \tag{3.4}$$

Assumption 3.4 ensures Z_kM_k to be positive semi-definite. The proof is exactly as for the PDPS, see, e.g., [7], but we include it for completeness.

Lemma 3.5. *Let $k \in \mathbb{N}$ and suppose Assumption 3.4 holds. Then*

$$Z_kM_k \geq \operatorname{diag}\left(0, 0, \varphi_k(1-\kappa)\operatorname{In}_X, \psi_{k+1}\varepsilon \operatorname{In}_Y\right) \geq 0 \quad \textit{for} \quad \varepsilon := 1 - \frac{\tau_k\sigma_k}{\kappa(1+2\tilde\gamma_{G^*}\sigma_k)}\|K\|^2 > 0.$$

Proof. By Young's inequality, for any $v = (u, w, x, y)$,

$$\begin{aligned} \langle Z_kM_kv|v\rangle &= \varphi_k\|x\|_X^2 + \psi_{k+1}\|y\|_Y^2 - 2\eta_k \left\langle x, K^\star y\right\rangle_X \\ &\geq \varphi_k(1-\kappa)\|x\|_X^2 + \psi_{k+1}\|y\|_Y^2 - \kappa^{-1}\varphi_k\tau_k^2\|K^\star y\|_X^2. \end{aligned}$$

Since $\varphi_k\tau_k^2 = \eta_k\tau_k = \psi_k\sigma_k\tau_k = \psi_{k+1}\sigma_k\tau_k/(1+2\tilde\gamma_{G^*}\sigma_k)$, the claim follows. □

3.3 GROWTH ESTIMATES AND MONOTONICITY

We start by deriving a three-point monotonicity estimate for H_k. This demands the somewhat strict bounds (3.5).

Lemma 3.6. *Let $k \in \mathbb{N}$. Suppose Assumptions 3.1, 3.3 and 3.4 hold and*

$$\gamma_F \geq \tilde{\gamma}_F + \varepsilon_u + \varepsilon_w + \frac{\lambda_{k+1}\pi_u + \theta_{k+1}\pi_w}{\eta_k}, \tag{3.5a}$$

$$\gamma_{G^*} \geq \tilde{\gamma}_{G^*}, \tag{3.5b}$$

$$\gamma_B \geq \frac{\lambda_{k+1}}{\lambda_k} + \frac{\theta_k}{\lambda_k}C_Q + \frac{\eta_k \mathcal{S}(\bar{w})}{4\varepsilon_w \lambda_k} + \frac{C_x \mu \eta_k}{2\lambda_k}, \quad \textit{and} \tag{3.5c}$$

$$\gamma_B \geq \frac{\theta_{k+1}}{\theta_k} + \frac{\eta_k \mathcal{S}(\bar{u})}{4\varepsilon_u \theta_k} + \frac{C_x \eta_k}{2\mu\theta_k} \tag{3.5d}$$

for some $\varepsilon_u, \varepsilon_w, \mu > 0$. Then H_k defined in (2.17) satisfies

$$\begin{aligned}\langle Z_k T_k H_k(v^{k+1}) | v^{k+1} - \bar{v}\rangle \geq{}& \frac{1}{2}\|v^{k+1} - \bar{v}\|^2_{Z_k\Xi_k} \\ &+ (\lambda_{k+1}\pi_u + \theta_{k+1}\pi_w)\|x^{k+1} - \bar{x}\|^2_X - (\lambda_k\pi_u + \theta_k\pi_w)\|x^k - \bar{x}\|^2_X \\ &+ \lambda_{k+1}\|u^{k+1} - \bar{u}\|^2_U - \lambda_k\|u^k - \bar{u}\|^2_U \\ &+ \theta_{k+1}\|w^{k+1} - \bar{w}\|^2_W - \theta_k\|w^k - \bar{w}\|^2_W.\end{aligned} \tag{3.6}$$

Proof. For brevity we denote $v = (u, w, x, y) := v^{k+1}$. Recall that $\bar{v} = (\bar{u}, \bar{w}, \bar{x}, \bar{y})$ satisfies by Assumption 3.1 the optimality conditions (2.9). Since Algorithm 2.1 guarantees the first two lines of H_k to be zero through the choice of M_k in (2.18), introducing $q_F := -\bar{\nabla}_x B(\bar{u}, \bar{w}) - K^\star \bar{y} \in \partial F(\bar{x})$ we expand

$$\begin{aligned}\langle Z_k T_k H_k(v) | v - \bar{v}\rangle &= \eta_k\langle \partial F(x) + \bar{\nabla}_x B(u, w) + K^\star y, x - \bar{x}\rangle_X + \eta_{k+1}\langle \partial G^*(y) - Kx, y - \bar{y}\rangle_Y \\ &= \eta_k\langle \partial F(x) - q_F, x - \bar{x}\rangle_X + \eta_k\langle \bar{\nabla}_x B(u, w) - \bar{\nabla}_x B(\bar{u}, \bar{w}), x - \bar{x}\rangle_X \\ &\quad + \eta_{k+1}\langle \partial G^*(y) - K\bar{x}, y - \bar{y}\rangle_Y + (\eta_k - \eta_{k+1})\langle K(x - \bar{x}), y - \bar{y}\rangle_Y.\end{aligned}$$

Using (3.4) we also have

$$\frac{1}{2}\|v - \bar{v}\|^2_{Z_k\Xi_k} = \eta_k\tilde{\gamma}_F\|x - \bar{x}\|^2_X + (\eta_k - \eta_{k+1})\,\langle K(x - \bar{x}), y - \bar{y}\rangle_Y + \eta_{k+1}\tilde{\gamma}_{G^*}\||y - \bar{y}\|^2_Y.$$

We now use the (strong) monotonicity of F and G^* with constants γ_F and γ_{G^*} contained Assumption 3.1 (i), as well as the splitting inequality Assumption 3.3 (iii). Thus

$$\begin{aligned}\langle Z_k T_k H_k(v) | v - \bar{v}\rangle \geq{}& \frac{1}{2}\|v - \bar{v}\|^2_{Z_k\Xi_k} + \eta_k(\gamma_F - \tilde{\gamma}_F)\|x - \bar{x}\|^2_X - (\lambda_k\pi_u + \theta_k\pi_w)\|x^k - \bar{x}\|^2_X \\ &+ \eta_{k+1}(\gamma_{G^*} - \tilde{\gamma}_{G^*})\|y - \bar{y}\|^2_Y + \eta_k\langle \bar{\nabla}_x B(u, w) - \bar{\nabla}_x B(\bar{u}, \bar{w}), x - \bar{x}\rangle_X \\ &+ (\lambda_k\gamma_B - \theta_k C_Q)\|u - \bar{u}\|^2_U - \lambda_k\|u^k - \bar{u}\|^2_U \\ &+ \theta_k\gamma_B\|w - \bar{w}\|^2_W - \theta_k\|w^k - \bar{w}\|^2_W.\end{aligned} \tag{3.7}$$

The Riesz equivalence (2.6), affine-linear-linear structure of B_x, Assumption 3.1 (iii) and (iv), and Young's inequality give

$$\begin{aligned}\eta_k\langle \bar{\nabla}_x B(u, w) - \bar{\nabla}_x B(\bar{u}, \bar{w}), x - \bar{x}\rangle_X &= \eta_k B_x(u, w, x - \bar{x}) - \eta_k B_x(\bar{u}, \bar{w}, x - \bar{x}) \\ &= \eta_k B_x(u, w, x - \bar{x}) + \eta_k B_x(\bar{u}, w - \bar{w}, x - \bar{x}) - \eta_k B_x(\bar{u}, w, x - \bar{x}) \\ &= \eta_k B_{xu}(u - \bar{u}, w - \bar{w}; x - \bar{x}) + \eta_k B_x(\bar{u}, w - \bar{w}; x - \bar{x}) + \eta_k B_{xu}(u - \bar{u}, \bar{w}; x - \bar{x}) \\ &\geq -\eta_k\left(\frac{\mathcal{S}(\bar{u})}{4\varepsilon_u} + \frac{C_x\mu}{2}\right)\|w - \bar{w}\|^2_W - \eta_k\left(\frac{\mathcal{S}(\bar{w})}{4\varepsilon_w} + \frac{C_x}{2\mu}\right)\|u - \bar{u}\|^2_U - \eta_k(\varepsilon_u + \varepsilon_w)\|x - \bar{x}\|^2_X\end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8), we obtain

$$
\begin{aligned}
\langle Z_k T_k H_k(v) | v - \bar v \rangle \geq{}& \frac{1}{2}\|v - \bar v\|^2_{Z_k \Xi_k} + \eta_{k+1}(\gamma_{G^*} - \tilde\gamma_{G^*})\|y - \bar y\|^2_Y \\
&+ \eta_k(\gamma_F - \tilde\gamma_F - \varepsilon_u - \varepsilon_w)\|x - \bar x\|^2_X - (\lambda_k \pi_u + \theta_k \pi_w)\|x^k - \bar x\|^2_X \\
&- \lambda_k \|u^k - \bar u\|^2_U + \lambda_k\left(\gamma_B - \frac{\theta_k}{\lambda_k} C_Q - \frac{\eta_k \mathcal{S}(\bar w)}{4\varepsilon_w \lambda_k} - \frac{C_x \mu \eta_k}{2\lambda_k}\right)\|u - \bar u\|^2_U \\
&- \theta_k \|w^k - \bar w\|^2_W + \theta_k\left(\gamma_B - \frac{\eta_k \mathcal{S}(\bar u)}{4\varepsilon_u \theta_k} - \frac{C_x \eta_k}{2\mu\theta_k}\right)\|w - \bar w\|^2_W.
\end{aligned}
$$

The claim now follows by applying (3.5). □

We now simplify and interpret (3.5).

Lemma 3.7. *Suppose $\gamma_F > \tilde\gamma_F > 0$ as well as $\gamma_{G^*} \geq \tilde\gamma_{G^*} \geq 0$ and that there exists $\omega, t > 0$ with $\omega\eta_{k+1} \leq \eta_k$ for all $k \in \mathbb{N}$, such that*

$$
\gamma_B \geq \omega^{-1} + tC_Q + \frac{2(1+t^{-1})}{\omega(\gamma_F - \tilde\gamma_F)^2}\left(\mathcal{S}(\bar u)\pi_w + t\mathcal{S}(\bar w)\pi_u + \frac{1}{2}\sqrt{t\pi_u\pi_w}C_x(\gamma_F - \tilde\gamma_F)\right). \tag{3.9}
$$

Then there exist $\varepsilon_u, \varepsilon_w, \mu > 0$ and, for all $k \in \mathbb{N}$, $\lambda_k, \theta_k > 0$ such that (3.5) holds. Moreover

$$
\lambda_k \pi_u + \theta_k \pi_w = \eta_k \omega \frac{\gamma_F - \tilde\gamma_F}{2}. \tag{3.10}
$$

Proof. We take

$$
\lambda_k := t^{-1} r \pi_u^{-1} \eta_k \quad\text{and}\quad \theta_k := r\pi_w^{-1}\eta_k \quad\text{for}\quad r := \frac{(\gamma_F - \tilde\gamma_F)\omega}{2(t^{-1}+1)} \quad\text{and}\quad c_k := \frac{\eta_{k+1}}{\eta_k}. \tag{3.11}
$$

These expressions readily give (3.10). We then take $\mu := (t\pi_u/\pi_w)^{-1/2}$,

$$
\varepsilon_u := \frac{\mathcal{S}(\bar u)}{\mathcal{S}(\bar u) + t\mathcal{S}(\bar w)} \frac{\gamma_F - \tilde\gamma_F}{2}, \quad\text{and}\quad \varepsilon_w := \frac{t\mathcal{S}(\bar w)}{\mathcal{S}(\bar u) + t\mathcal{S}(\bar w)} \frac{\gamma_F - \tilde\gamma_F}{2}.
$$

Since both

$$
\frac{\lambda_{k+1}\pi_u + \theta_{k+1}\pi_w}{\eta_k} = c_k r(t^{-1}+1) = c_k \omega \frac{\gamma_F - \tilde\gamma_F}{2} \leq \frac{\gamma_F - \tilde\gamma_F}{2}
$$

and $\varepsilon_u + \varepsilon_w = (\gamma_F - \tilde\gamma_F)/2$, (3.5a) is readily verified, while (3.5b) we have assumed. Inserting $\lambda_k, \theta_k, \eta_k$, and μ, we also rewrite (3.5c) and (3.5d) as

$$
\gamma_B \geq c_k + tC_Q + \frac{t\mathcal{S}(\bar w)\pi_u}{4\varepsilon_w r} + \frac{\sqrt{t\pi_u\pi_w}C_x}{2r} \quad\text{and}\quad \gamma_B \geq c_k + \frac{\mathcal{S}(\bar u)\pi_w}{4\varepsilon_u r} + \frac{\sqrt{t\pi_u\pi_w}C_x}{2r}.
$$

After also inserting $\varepsilon_u, \varepsilon_w$, and r, and using $\omega c_k \leq 1$, these are readily verified by (3.9). □

Remark 3.8. Since $\eta_{k+1} \geq \eta_k$ for convergent algorithms, i.e., $\omega^{-1} \geq 1$, letting $\omega = 1$ and $\tilde\gamma_F = 0$ in (3.9), we obtain at the solution $(\bar u, \bar w, \bar x, \bar y)$ a fundamental "second order growth" and splitting condition (via C_Q, π_u, and π_w) that cannot be avoided by step length parameter choices.

Our convergence proof is based based on the next Féjer-type monotonicity estimate with respect to the iteration-dependent norms $\|\cdot\|_{Z_k\tilde M_k}$. Here $\tilde M_k \in \mathbb{L}(U \times W \times X \times Y; U^* \times W^* \times X \times Y)$ modifies M_k defined in (2.18) as

$$
\tilde M_k := M_k + \operatorname{diag}\begin{pmatrix}\mathrm{In}_U & \mathrm{In}_W & \varphi_k^{-1}(\lambda_k\pi_u + \theta_k\pi_w)\,\mathrm{Id}_X & 0\end{pmatrix}. \tag{3.12}
$$

By (3.2) and Assumption 3.4, this satisfies

$$
Z_k\tilde{M}_k = \begin{pmatrix} \lambda_k \operatorname{In}_U & & & \\ & \theta_k \operatorname{In}_W & & \\ & & (\varphi_k + \lambda_k\pi_u + \theta_k\pi_w)\operatorname{In}_X & -\eta_k \operatorname{In}_X K^\star \\ & & -\eta_k \operatorname{In}_Y K & \psi_{k+1}\operatorname{In}_Y \end{pmatrix}. \tag{3.13}
$$

Lemma 3.9. *Suppose Assumptions 3.1 and 3.4 hold as does Assumption 3.3 and (3.5) for $k = 0, \ldots, N$. Given v^0, let $v^1, \ldots, v^{N-1}$ be produced by Algorithm 2.1. Then*

$$
\frac{1}{2}\|v^{k+1} - \bar{v}\|^2_{Z_{k+1}\tilde{M}_{k+1}} + \frac{1}{2}\|v^{k+1} - v^k\|^2_{Z_kM_k} \le \frac{1}{2}\|v^k - \bar{v}\|^2_{Z_k\tilde{M}_k} \quad (k = 0, \ldots, N-1) \tag{3.14}
$$

where all the terms are non-negative.

Proof. Lemma 3.6 gives the estimate

$$
\begin{aligned}
\langle Z_kT_kH_k(v^{k+1}) | v^{k+1} - \bar{v}\rangle \ge{}& \frac{1}{2}\|v^{k+1} - \bar{v}\|^2_{Z_k\Xi_k} \\
&+ (\lambda_{k+1}\pi_u + \theta_{k+1}\pi_w)\|x^{k+1} - \bar{x}\|^2_X - (\lambda_k\pi_u + \theta_k\pi_w)\|x^k - \bar{x}\|^2_X \\
&+ \lambda_{k+1}\|u^{k+1} - \bar{u}\|^2_U - \lambda_k\|u^k - \bar{u}\|^2_U \\
&+ \theta_{k+1}\|w^{k+1} - \bar{w}\|^2_W - \theta_k\|w^k - \bar{w}\|^2_W \\
={}& \frac{1}{2}\|v^{k+1} - \bar{v}\|^2_{Z_{k+1}(\tilde{M}_{k+1} - M_{k+1}) + Z_k\Xi_k} - \frac{1}{2}\|v^k - \bar{v}\|^2_{Z_k(\tilde{M}_k - M_k)}.
\end{aligned} \tag{3.15}
$$

By the implicit form (2.19) of Algorithm 2.1, we have $-Z_kM_k(v^{k+1} - v^k) \in Z_kT_kH_k(v^{k+1})$. Thus (3.15) combined with the three-point identity (1.2) for the operator $M = Z_kM_k$ yields

$$
\frac{1}{2}\|v^k - \bar{v}\|^2_{Z_k\tilde{M}_k} \ge \frac{1}{2}\|v^{k+1} - \bar{v}\|^2_{Z_{k+1}(\tilde{M}_{k+1} - M_{k+1}) + Z_k(M_k + \Xi_k)} + \frac{1}{2}\|v^{k+1} - v^k\|^2_{Z_kM_k}
$$

Therefore (3.14) follows by applying (3.3), i.e., $Z_k(M_k + \Xi_k) = Z_{k+1}M_{k+1} + D_k$, where the skew symmetric term D_k does not contribute to the norms. Finally, we have $Z_k\tilde{M}_k \ge Z_kM_k \ge 0$ by Lemma 3.5, proving the non-negativity of all the terms. □

3.4 MAIN RESULTS

We can now state our main convergence theorems. In terms of assumptions, the only fundamental difference between the accelerated $O(1/N)$ and the linear convergence result is that the latter requires G^* to be strongly convex and the former doesn't. Both require sufficient second order growth in terms of the respective technical conditions (3.16b) or (3.19b). The step length parameters differ.

Theorem 3.10 (Accelerated convergence). *Suppose Assumptions 3.1 and 3.3 hold with $\gamma_F > 0$. Put $\tilde{\gamma}_{G^*} = 0$ and pick $\tau_0, \sigma_0, \kappa, t > 0$ and $0 < \tilde{\gamma}_F < \gamma_F$ satisfying*

$$
1 > \kappa \ge \tau_0\sigma_0\|K\|^2 \quad \text{and} \tag{3.16a}
$$

$$
\gamma_B \ge \omega_0^{-1} + tC_Q + \frac{2(1 + t^{-1})}{\omega_0(\gamma_F - \tilde{\gamma}_F)^2}\left(\mathcal{S}(\bar{u})\pi_w + t\mathcal{S}(\bar{w})\pi_u + \frac{1}{2}\sqrt{t\pi_u\pi_w}C_x(\gamma_F - \tilde{\gamma}_F)\right), \tag{3.16b}
$$

where ω_0 is defined as part of the update rules

$$
\tau_{k+1} := \tau_k\omega_k, \quad \sigma_{k+1} := \sigma_k/\omega_k, \quad \text{and} \quad \omega_k := 1/\sqrt{1 + 2\tilde{\gamma}_F\tau_k} \quad (k \in \mathbb{N}).
$$

Let $\{v^{k+1}\}_{k\in\mathbb{N}}$ be generated by Algorithm 2.1 for any $v^0 \in U \times W \times X \times Y$. Then $x^k \to \bar{x}$ in X; $u^k \to \bar{u}$ in U; and $w^k \to \bar{w}$ in W, all strongly at the rate $O(1/N)$.

Proof. We use Lemma 3.9, whose assumptions we now verify. Assumptions 3.1 and 3.3 we have assumed. As shown in [34, 7], Assumption 3.4 holds with $\psi_k \equiv \sigma_0^{-1}\tau_0, \varphi_0 = 1$, and $\varphi_{k+1} := \varphi_k/\omega_k^2$. Moreover, $\{\varphi_k\}_{k\in\mathbb{N}}$ grows at the rate $\Omega(k^2)$. Hence

$$\eta_{k+1} = \omega_k^{-1}\eta_k = \sqrt{1+2\tilde\gamma_F\tau_k}\eta_k \le \omega_0^{-1}\eta_k \quad \text{for} \quad \omega_0^{-1} = \sqrt{1+2\tilde\gamma_F\tau_0}.$$

Thus (3.16) verifies (3.9) so that Lemma 3.7 verifies (3.5). Thus we may apply Lemma 3.9. By summing its result over $k = 0, \ldots, N-1$, we get

$$\frac{1}{2}\|v^N - \bar v\|^2_{Z_N\tilde M_N} \le \frac{1}{2}\|v^0 - \bar v\|^2_{Z_0\tilde M_0}. \tag{3.17}$$

By (3.2), (3.13), and Lemma 3.5 we have

$$Z_k\tilde{M_k} \ge Z_k M_k \ge \operatorname{diag}\left(\lambda_k \operatorname{In}_U \quad \theta_k \operatorname{In}_W \quad \varphi_k(1-\kappa)\operatorname{In}_X \quad \psi_{k+1}\varepsilon \operatorname{In}_Y\right) \ge 0. \tag{3.18}$$

where $\varepsilon := 1 - \tau_k\sigma_k\kappa^{-1}\|K\|^2 = 1 - \tau_0\sigma_0\kappa^{-1}\|K\|^2 > 0$ by assumption. By Lemma 3.7, $\{\lambda_k\}_{k\in\mathbb{N}}$ and $\{\theta_k\}_{k\in\mathbb{N}}$ grow at the same $\Omega(k^2)$ rate as $\{\varphi_k\}_{k\in\mathbb{N}}$. Therefore (3.17) and (3.18) establish $\|x^k - \bar x\|_X^2 \to 0$ as well as $\|u^k - \bar u\|_U^2$ and $\|w^k - \bar w\|_W^2 \to 0$, all at the rate $O(1/N^2)$. The claim follows by removing the squares. □

Theorem 3.11 (Linear convergence). *Suppose Assumptions 3.1 and 3.3 hold with both $\gamma_F > 0$ and $\gamma_{G^*} > 0$. Pick $\tau, \kappa, t > 0$, $0 < \tilde\gamma_F \le \gamma_F$, $0 < \tilde\gamma_{G^*} \le \gamma_{G^*}$ satisfying*

$$1 > \kappa \ge \tau^2\tilde\gamma_{G^*}^{-1}\tilde\gamma_F\|K\|^2 \quad \textit{and} \tag{3.19a}$$

$$\gamma_B \ge \omega^{-1} + tC_Q + \frac{2(1+t^{-1})}{\omega(\gamma_F-\tilde\gamma_F)^2}\left(\mathcal{S}(\bar u)\pi_w + t\mathcal{S}(\bar w)\pi_u + \frac{1}{2}\sqrt{t\pi_u\pi_w}C_x(\gamma_F - \tilde\gamma_F)\right) \tag{3.19b}$$

for

$$\sigma := \tilde\gamma_{G^*}^{-1}\tilde\gamma_F\tau \quad \textit{and} \quad \omega := 1/(1+2\tilde\gamma_F\tau) = 1/(1+\tilde\gamma_{G^*}\sigma).$$

Take $\tau_k \equiv \tau$, $\sigma_k \equiv \sigma$, and $\omega_k \equiv \omega$. Let $\{v^{k+1}\}_{k\in\mathbb{N}}$ be generated by Algorithm 2.1 for any $v^0 \in U \times W \times X \times Y$. Then $x^k \to \bar x$ in X; $u^k \to \bar u$ in U; and $w^k \to \bar w$ in W, all strongly at a linear rate.

Proof. As shown in [34, 7], Assumption 3.4 is satisfied for $\varphi_0 = 1$, $\psi_0 = \sigma^{-1}\tau$, $\varphi_{k+1} := \varphi_k/\omega_k$, and $\psi_{k+1} := \psi_k/\omega_k$. Moreover, both $\{\varphi_k\}_{k\in\mathbb{N}}$ and $\{\psi_k\}_{k\in\mathbb{N}}$ grow exponentially and $\eta_{k+1} \le \omega^{-1}\eta_k$. Thus (3.19) verifies (3.9) with $c = \omega^{-1}$ so that Lemma 3.7 verifies (3.5). The rest follows as in the proof of Theorem 3.10. □

Theorems 3.10 and 3.11 show global convergence, but may require a very constricted dom F through the constant C_x in Assumption 3.1 (iv). In Appendix B we relax the constant by localizing the convergence.

Remark 3.12 (Linear and sufficiently linear PDEs). For linear PDEs, i.e., when B_x does not depend on u, we have $C_x = 0$ and $\mathcal{S}(\bar w) = 0$, as observed in Remark 3.2. Moreover, for typical solvers for the adjoint PDE, we would have $\pi_w = 0$, as B_u does not then depend on x. In that case, by taking $t \searrow 0$, (3.16b) (and likewise (3.19b)) reduces to $\gamma_B > \omega_0^{-1}$. Practically this means that the convergence rate factor ω_0^{-1} has to be bounded by the inverse contractivity factor γ_B of the linear system solver. If $\gamma_B > 1$, as we should have, this condition can be satisfied by suitable choices of $\tilde\gamma_F \in (0, \gamma_F]$ and $\tilde\gamma_{G^*}$. By extension then, the conditions (3.16b) and (3.19b) are satisfiable for small t when the PDE is "sufficiently linear".

Remark 3.13 (Weak convergence). It is possible to prove weak convergence when $\omega \equiv 1$ and $\tau \equiv \tau_0$, $\sigma \equiv \sigma_0$ satisfy (3.16). The proof is based on an extension of Opial's lemma to the quantitative Féjer monotonicity (3.14). We have not included the proof since it is technical, and does not permit reducing assumptions from those of Theorems 3.10 and 3.11. We refer to [4] for the corresponding proof for the NL-PDPS.

4 SPLITTINGS AND PARTIAL DIFFERENTIAL EQUATIONS

We now prove Assumption 3.1 and derive explicit expressions for the operator $\bar{\nabla}_x B$ from (2.6). We do this in Section 4.1 for some sample PDEs. Then in Section 4.2 we study the satisfaction of Assumption 3.3 for Gauss–Seidel and Jacobi splitting, as well as a simple infinite-dimensional example without splitting. We briefly discuss a quasi-conjugate gradient splitting to illustrate the generality of our approach. We conclude with a discussion of the convergence theory and discretisation in Section 4.3.

4.1 PARTIAL DIFFERENTIAL EQUATIONS AND RIESZ REPRESENTATIONS

Let $\mathrm{Sym}^d \subset \mathbb{R}^{d\times d}$ stand for the symmetric matrices. Recall that in Example 2.2, to ensure the continuity of B, we needed in practise that at least one of the spaces U, W, or X be finite-dimensional. The same will be the case here. Accordingly, with $\Omega \subset \mathbb{R}^d$ a Lipschitz domain, we take

$$x = (A, c) \in X := X_1 \times X_2 \quad \text{for subspaces} \quad X_1 \subset L^2(\Omega; \mathrm{Sym}^d) \quad \text{and} \quad X_2 \subset L^2(\Omega), \tag{4.1a}$$

as well as $U \subset H^1(\Omega)$ and $W \subset H^1_0(\Omega) \times H^{1/2}(\partial\Omega)$ such that

$$B(u, w; x) := B_x(u, w; x) + B_{\mathrm{const}}(u, w) \quad \text{for} \quad u \in U,\ w \in W,\ x \in X \tag{4.1b}$$

is continuous, where, writing $w = (w_\Omega, w_\partial)$,

$$B_x(u, w; x) := \langle \nabla u, A\nabla w_\Omega \rangle_{L^2(\Omega)} + \langle cu, w_\Omega \rangle_{L^2(\Omega)} \quad \text{and} \tag{4.1c}$$

$$B_{\mathrm{const}}(u, w) := \langle \mathrm{trace}_{\partial\Omega}\, u, w_\partial \rangle_{L^2(\partial\Omega)}. \tag{4.1d}$$

Thus B_{const} models the nonhomogeneous Dirichlet boundary condition $u = g$ on $\partial\Omega$ for some $g \in H^{-\frac{1}{2}}(\partial\Omega)$. Correspondingly we take for some $L_0 \in H^{-1}(\Omega)$ the right-hand-side

$$Lw := L_0 w_\Omega + \langle g, w_\partial \rangle_{L^2(\partial\Omega)}. \tag{4.1e}$$

The next lemma verifies the PDE components of Assumption 3.1. Afterwards we look at particular choices of X_1 and X_2. We could also take $W = H^1(\Omega)$, $w = w_\Omega$, $L = L_0$, and $B_{\mathrm{const}} = 0$ to model Neumann boundary conditions, and the result would still hold. In the range spaces of $L^p(\Omega; \mathbb{R}^d)$, $W^{1,p}(\Omega)$, and $L^p(\Omega; \mathbb{R}^{d\times d})$, we use the Euclidean norm in $\mathbb{R}^d$ and the spectral norm $\|\cdot\|_2$ in $\mathbb{R}^{d\times d}$.

Lemma 4.1. *Assume* (4.1) *and that* $\mathrm{dom}\, F \subset L^\infty(\Omega; \mathbb{R}^{d\times d}) \times L^\infty(\Omega)$*. Then:*

(ii′) Assumption 3.1 (ii) holds if there exists $\lambda \in (0, 1)$ *such that*

$$A(\xi) \geq \lambda\, \mathrm{Id} \quad \textit{and} \quad |c(\xi)| \geq \lambda \quad \textit{for all} \quad \xi \in \Omega \quad \textit{and} \quad (A, c) \in (X_1 \times X_2) \cap \mathrm{dom}\, F.$$

Suppose then that (2.9) *is solved by* $\bar{v} = (\bar{u}, \bar{w}, \bar{x}, \bar{y})$ *with* $\bar{x} = (\bar{A}, \bar{c}) \in \mathrm{dom}\, F \subset (X_1 \times X_2)$, $\bar{u} \in H^1(\Omega)$, $\bar{w} = (\bar{w}_\Omega, \bar{w}_\partial) \in H^1_0(\Omega) \times H^{1/2}(\partial\Omega)$*. If* $\|\bar{u}\|_{W^{1,\infty}(\Omega)}, \|\bar{w}\|_{W^{1,\infty}(\Omega)} < \infty$*, and* $\bar{y} \in Y$ *for a Hilbert space* Y*, then also:*

(iii′) Assumption 3.1 (iii) holds with $\mathcal{S}(\bar{u}) = \|\bar{u}\|^2_{W^{1,\infty}(\Omega)}$ *and* $\mathcal{S}(\bar{w}) = \|\bar{w}_\Omega\|^2_{W^{1,\infty}(\Omega)}$.

(iv′) Assumption 3.1 (iv) holds with

$$C_x = \sup_{(A,c)\in\mathrm{dom}\, F} \|A - \bar{A}\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} + \|c - \bar{c}\|_{L^\infty(\Omega)}.$$

Remark 4.2. On bounded Ω the condition $\|\bar{u}\|_{W^{1,\infty}(\Omega)} < \infty$ is stronger than $\bar{u} \in H^1(\Omega)$. We include both to emphasise that the latter defines the Hilbert space structure and topology that we generally work with, while the former is a technical restriction that arises from our proofs. Under appropriate smoothness conditions on $\bar{x}$, the boundary of Ω, as well as the boundary data, standard elliptic theory proves that $\bar{u} \in H^1(\Omega)$ is a classical solution, hence Lipschitz and $W^{1,\infty}(\Omega)$ on the whole domain; see, e.g., [9].

Proof. For (ii′), we identify $g \in H^{-1/2}(\partial\Omega)$ with $\hat{g} \in H^{1/2}(\partial\Omega)$ by the Riesz mapping and fix $\hat{u} \in H^1(\Omega)$ with $\operatorname{trace}_{\partial\Omega} \hat{u} = \hat{g}$. This is possible by the definition of $H^{1/2}(\partial\Omega)$. By the Lax–Milgram lemma there is then a unique solution $v \in H^1_0(\Omega)$ to

$$\langle \nabla v, A\nabla w_\Omega\rangle_{L^2(\Omega)} + \langle cv, w_\Omega\rangle_{L^2(\Omega)} = L_0 w_\Omega - B_x(\hat{u}, w_\Omega; x) \quad \text{for all} \quad w_\Omega \in H^1_0(\Omega),$$

Now $u = v + \hat{u}$ satisfies $B(u, w; x) = Lw$ and is independent of the choice of $\hat{u}$. Analogously we prove the existence of a solution to the adjoint equation.

To prove (iv′), pick arbitrary $u \in H^1(\Omega)$, $w = (w_\Omega, w_\partial) \in H^1_0(\Omega) \times H^{1/2}(\partial\Omega)$, and $x = (A, c) \in (X_1 \times X_2) \cap \operatorname{dom} F$. Hölder's inequality and the symmetry of $A(\xi)$ give

$$\begin{aligned}\langle \nabla u, A\nabla w_\Omega\rangle_{L^2(\Omega)} &\le \|\nabla w_\Omega\|_{L^2(\Omega;\mathbb{R}^d)} \left(\int_\Omega \|A(\xi)\nabla u(\xi)\|_2^2 \, d\xi\right)^{1/2} \\ &\le \|\nabla w_\Omega\|_{L^2(\Omega;\mathbb{R}^d)} \|A\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \|\nabla u\|_{L^2(\Omega)}.\end{aligned}$$

Therefore, as claimed

$$\begin{aligned}B_x(u, w; x - \bar{x}) &\le \|A - \bar{A}\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} \|\nabla u\|_{L^2(\Omega;\mathbb{R}^d)} \|\nabla w_\Omega\|_{L^2(\Omega;\mathbb{R}^d)} \\ &\quad + \|c - \bar{c}\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|w_\Omega\|_{L^2(\Omega)} \\ &\le \big(\|A - \bar{A}\|_{L^\infty(\Omega;\mathbb{R}^{d\times d})} + \|c - \bar{c}\|_{L^\infty(\Omega)}\big) \|u\|_{H^1(\Omega)} \|w_\Omega\|_{H^1(\Omega)} \\ &\le C_x \|u\|_{H^1(\Omega)} \|w_\Omega\|_{H^1(\Omega)}.\end{aligned}$$

For (iii′), using Hölder's twice inequality and the symmetry of $A(\xi)$, we estimate

$$\begin{aligned}\langle \nabla u, A\nabla w_\Omega\rangle_{L^2(\Omega)} &\le \|\nabla w_\Omega\|_{L^\infty(\Omega;\mathbb{R}^d)} \int_\Omega \|A(\xi)\nabla u(\xi)\|_2 \, d\xi \\ &\le \|\nabla w_\Omega\|_{L^\infty(\Omega;\mathbb{R}^d)} \|A\|_{L^2(\Omega;\mathbb{R}^{d\times d})} \|\nabla u\|_{L^2(\Omega)}.\end{aligned}$$

Hence

$$\begin{aligned}B_x(u, \bar{w}; x) &\le \|A\|_{L^2(\Omega;\mathbb{R}^{d\times d})} \|\nabla u\|_{L^2(\Omega;\mathbb{R}^d)} \|\nabla \bar{w}_\Omega\|_{L^\infty(\Omega;\mathbb{R}^d)} \\ &\quad + \|c\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} \|\bar{w}_\Omega\|_{L^\infty(\Omega)} \\ &\le \big(\|\nabla \bar{w}_\Omega\|_{L^\infty(\Omega;\mathbb{R}^d)} + \|\bar{w}_\Omega\|_{L^\infty(\Omega)}\big) \|u\|_{H^1(\Omega)} \big(\|A\|_{L^2(\Omega;\mathbb{R}^{d\times d})} + \|c\|_{L^2}\big) \\ &= \|\bar{w}_\Omega\|_{W^{1,\infty}(\Omega)} \|u\|_{H^1(\Omega)} \|x\|_X.\end{aligned}$$

Thus we may take as claimed $\mathcal{S}(\bar{w}) = \|\bar{w}_\Omega\|^2_{W^{1,\infty}(\Omega)}$, and analogously $\mathcal{S}(\bar{u}) = \|\bar{u}\|^2_{W^{1,\infty}(\Omega)}$. □

To describe $\bar{\nabla}_x B$ we denote the double dot product and the outer product by

$$A : \tilde{A} = \sum_{ij} A_{ij}\tilde{A}_{ij}, \quad \text{and} \quad v \otimes w = vw^T \quad \text{for} \quad A, \tilde{A} \in \mathbb{R}^{d\times d} \quad \text{and} \quad v, w \in \mathbb{R}^d$$

Observe the identity $v^T A w = A : (v \otimes w)$.

Example 4.3 (General case). In the fully general case, formally and without regard for the solvability of the PDE (2.2), we equip $X_1 = L^2(\Omega; \mathbb{R}^{d\times d})$ with the inner product $\langle A_1, A_2\rangle_{X_1} := \int_\Omega A_1(\xi) : A_2(\xi)\, d\xi$ and $X_2 = L^2(\Omega; \mathbb{R})$ with the standard inner product in $L^2(\Omega; \mathbb{R})$. Then for all $u \in U$, $w \in W$, and $(d, h) \in X_1 \times X_2$, we have

$$B_x(u, w; (d, h)) = \langle \nabla u, d\nabla w\rangle_{L^2(\Omega)} + \langle hu, w\rangle_{L^2(\Omega)} = \langle \nabla u \otimes \nabla w, d\rangle_{X_1} + \langle uw, h\rangle_{X_2}.$$

Therefore the Riesz representation $\bar{\nabla}_x B$ has pointwise in Ω the expression

$$\bar{\nabla}_x B(u, w) = \begin{pmatrix} \nabla u \otimes \nabla w \\ uw \end{pmatrix}.$$

The constant C_x is as provided by Lemma 4.1.

Example 4.4 (Scalar function diffusion coefficient). Let then $X_1 := \{\xi \mapsto a(\xi)\,\mathrm{Id} \mid a \in L^2(\Omega)\}$. X_1 is isometrically isomorphic with $L^2(\Omega)$ since the spectral norm $\|a(\xi)\,\mathrm{Id}\,\|_2 = |a(\xi)|$. We may therefore identify X_1 and $L^2(\Omega)$. We also observe that the term $\langle \nabla u, A\nabla w\rangle_{L^2(\Omega)} = \langle a, \nabla u\cdot\nabla w\rangle_{X_1}$. Hence, pointwise in Ω,

$$\bar{\nabla}_x B(u,w) = \begin{pmatrix} \nabla u \cdot \nabla w \\ uw \end{pmatrix}.$$

According to Lemma 4.1, the constant

$$C_x = \sup_{(a,c)\in \operatorname{dom} F} \|a - \bar{a}\|_{L^\infty(\Omega)} + \|c - \bar{c}\|_{L^\infty(\Omega)}.$$

Example 4.5 (Spatially uniform coefficients). Let $X_1 := \{\xi \mapsto \tilde{A} \mid \tilde{A} \in \mathrm{Sym}^d\} \subset L^2(\Omega; \mathrm{Sym}^d)$ and $X_2 := \{\xi \mapsto \tilde{c} \mid \tilde{c} \in \mathbb{R}\} \subset L^2(\Omega)$ consist of constant functions $A : \xi \mapsto \tilde{A}$ and $c : \xi \mapsto \tilde{c}$ on the bounded domain Ω. Then $\|x\|_{X_1\times X_2} = |\Omega|^{1/2}(\|\tilde{A}\|_2 + |\tilde{c}|)$ for all $x = (A,c) \in X_1 \times X_2$. We may thus identify X_1 and X_2 with $\mathbb{R}^{d\times d}$ and $\mathbb{R}$ if we weigh the norms by $|\Omega|^{1/2}$. We have

$$\langle \nabla u, A\nabla w\rangle_{L^2(\Omega)} = \int_\Omega \tilde{A} : \nabla u \otimes \nabla w \, d\xi = \tilde{A} : \int_\Omega \nabla u \otimes \nabla w \, d\xi.$$

Thus

$$\bar{\nabla}_x B(u,w) = \begin{pmatrix} \int_\Omega \nabla u \otimes \nabla w \, d\xi \\ \int_\Omega uw \, d\xi \end{pmatrix}.$$

According to Lemma 4.1, the constant

$$C_x = \sup_{(A,c)\in \operatorname{dom} F} \|\tilde{A} - \bar{\tilde{A}}\|_2 + |\tilde{c} - \bar{\tilde{c}}|.$$

4.2 SPLITTINGS

We now discuss linear system splittings and Assumption 3.3. Throughout this subsection we assume that

$$B(u,w;x) = \langle A_x u + f_x | w\rangle \quad\text{and}\quad Lw = \langle b|w\rangle \tag{4.2}$$

with $A_x \in \mathbb{L}(U; W^*)$ invertible for $x \in X$, and $f_x, b \in W^*$. Then for fixed $x \in X$ the weak PDE (2.2) and the adjoint $B_u(\,\cdot\,, w, x) = -Q'(u)$ reduce to the linear equations

$$A_x u = b - f_x \quad\text{and}\quad A_x^* w = -Q'(u),$$

where $A_x^* \in \mathbb{L}(W; U^*)$ is the dual product adjoint of A_x restricted to $W \hookrightarrow W^{**}$.

The basic splittings The next lemma helps to prove Assumption 3.3 subject to a control on the rate of dependence of A on x. In its setting, with $A_x = N_x + M_x$ with N_x "easily" invertible, Lines 3 and 4 of Algorithm 2.1 are given by (2.15).

Theorem 4.6. *In the setting* (4.2)*, suppose* Assumption 3.1 *holds and*

$$\|A_x - A_{\tilde{x}}\|_{\mathbb{L}(U;W^*)} \le L_A \|x - \tilde{x}\|_X \quad \textit{and} \quad \|f_x - f_{\tilde{x}}\|_{W^*} \le L_f \|x - \tilde{x}\|_X \quad (x, \tilde{x} \in \operatorname{dom} F) \tag{4.3}$$

for some $L_A \ge 0$. Split $A_x = N_x + M_x$ with N_x invertible, and assume there exist $\alpha \in [0,1)$ and $\gamma_N > 0$ such that all

$$\|N_x^{-1} M_x\|_{\mathbb{L}(U;U)}, \|N_x^{-1,*} M_x^*\|_{\mathbb{L}(W;W)} \le \alpha \quad \textit{and} \quad \gamma_N \|N_x^{-1}\|_{\mathbb{L}(W^*;U)} \le 1 \quad (x \in \operatorname{dom} F). \tag{4.4}$$

Also suppose ∇Q is L_Q-Lipschitz. For any $\gamma_B \in (1, 1/\alpha^2)$, $\lambda \in (0,1)$, and $\beta > 0$, set

$$\pi_w = \left(1+\beta+\frac{\alpha^2\gamma_B}{\lambda(1-\alpha^2\gamma_B)}\right)\frac{\gamma_B L_A^2\|\bar w\|_W^2}{\gamma_N^2}, \qquad C_Q = \left(\frac{1+\beta}{\beta}+\frac{\alpha^2\gamma_B}{(1-\lambda)(1-\alpha^2\gamma_B)}\right)\frac{\gamma_B L_Q^2}{\gamma_N^2}, \quad \textit{and}$$
$$\pi_u = \left(1+\beta+\frac{\alpha^2\gamma_B}{\lambda(1-\alpha^2\gamma_B)}\right)\frac{\gamma_B L_A^2\|\bar u\|_U^2}{\gamma_N^2} + \left(\frac{1+\beta}{\beta}+\frac{\alpha^2\gamma_B}{(1-\lambda)(1-\alpha^2\gamma_B)}\right)\frac{\gamma_B L_f^2}{\gamma_N^2}.$$

Let $\Gamma_k(u,w,x) = \langle M_x u|w\rangle$ and $\Upsilon_k(u,w,x) = \langle u|M_x^* w\rangle$. Then Assumption 3.3 holds for all $k \in \mathbb{N}$ with $\{v^{k+1}\}_{k=0}^\infty$ generated by Algorithm 2.1 for any $v^0 \in U \times W \times X \times Y$.

Proof. Assumption 3.3 (i) holds by construction, and (ii) by the assumed invertibility of N_x for $x \in \operatorname{dom} F$. We only consider the second inequality of (iii) for Υ, the proof of the first inequality for Γ being analogous with $-Q'(u)$ replaced by $b - f_x$. We thus need to prove

$$\|w^k - \bar w\|_W^2 \geq \gamma_B\|w^{k+1}-\bar w\|_W^2 - C_Q\|u^{k+1}-\bar u\|_U^2 - \pi_B\|x^k - \bar x\|_X^2. \tag{4.5}$$

Using (2.15) with $A_{\bar x}^*\bar w = -Q'(\bar u)$ and $A_{x^k}^*\bar w = N_{x^k}^*\bar w + M_{x^k}^*\bar w$, we expand

$$\begin{aligned} w^{k+1} - \bar w &= N_{x^k}^{-1,*}(-Q'(u^{k+1}) - M_{x^k}^* w^k) - \bar w \\ &= N_{x^k}^{-1,*}[Q'(\bar u) - Q'(u^{k+1})] + N_{x^k}^{-1,*}(A_{\bar x}^* - A_{x^k}^*)\bar w - N_{x^k}^{-1,*}M_{x^k}^*(w^k - \bar w). \end{aligned}$$

Expanding $\|w^{k+1}-\bar w\|_W^2$ and applying the triangle inequality, and Young's inequality thrice, yields

$$\begin{aligned} \|w^{k+1}-\bar w\|_W^2 \leq{}& \left(1+\frac{\alpha^2\gamma_B}{\lambda(1-\alpha^2\gamma_B)}+\beta\right)\|N_{x^k}^{-1,*}(A_{\bar x}^* - A_{x^k}^*)\bar w\|_W^2 + \frac{1}{\alpha^2\gamma_B}\|N_{x^k}^{-1,*}M_{x^k}^*(w^k-\bar w)\|_W^2 \\ &+ \left(\frac{1+\beta}{\beta}+\frac{\alpha^2\gamma_B}{(1-\lambda)(1-\alpha^2\gamma_B)}\right)\|N_{x^k}^{-1,*}[Q'(u^{k+1}) - Q'(\bar u)]\|_W^2. \end{aligned}$$

Note that the first part of (4.3) and the second part (4.4) hold also for the adjoints A_x^* and N_x^* in the corresponding spaces. Therefore, we establish $\|N_{x^k}^{-1,*}(A_{\bar x}^* - A_{x^k}^*)\bar w\|_W^2 \leq \gamma_N^{-2}L_A^2\|\bar w\|_W^2\|\bar x - x^k\|_X^2$, $\|N_{x^k}^{-1,*}[Q'(u^{k+1}) - Q'(\bar u)]\|_W^2 \leq \gamma_N^{-2}L_Q^2\|u^{k+1}-\bar u\|_X^2$, and $\|N_{x^k}^{-1,*}M_{x^k}^*(w^k - \bar w)\|_W^2 \leq \alpha^2\gamma_B\|w^k - \bar w\|_W^2$. Taking π_w and C_Q as stated, we therefore obtain (4.5). □

For our first, infinite-dimensional example of the satisfaction of the conditions of Theorem 4.6, and hence of Assumption 3.3, note that we have in general

$$\|N_x^{-1}\|_{\mathbb{L}(W^*;U)} = \sup_{w^*}\frac{\|N_x^{-1}w^*\|_U}{\|w^*\|_{W^*}} = \sup_u \frac{\|u\|_U}{\|N_x u\|_{W^*}} = \sup_u \inf_w \frac{\|u\|_U\|w\|_w}{\langle N_x u|w\rangle}$$

and

$$\|A_x - A_{\bar x}\|_{\mathbb{L}(U;W^*)} = \sup_u \frac{\|[A_x - A_{\bar x}]u\|_{W^*}}{\|u\|_U} = \sup_{u,w}\frac{\langle [A_x - A_{\bar x}]u|w\rangle}{\|u\|_U\|w\|_W}.$$

Example 4.7 (No splitting of a weighted Laplacian in H^1). Let $U = W = H^1(\Omega)$, $X = \mathbb{R}$, and $N_x = A_x = x\nabla^*\nabla \in \mathbb{L}(H^1(\Omega); H^1(\Omega)^*)$ be the Laplacian weighted by $x \in (0,\infty)$. Then

$$\|N_x^{-1}\|_{\mathbb{L}(W^*;U)} = \sup_u \inf_w \frac{\|u\|_{H^1(\Omega)}^2}{x\langle\nabla u, \nabla w\rangle_{L^2(\Omega)}} \leq \sup_u \frac{\|u\|_{H^1(\Omega)}^2}{x\|\nabla u\|_{L^2(\Omega)}^2}.$$

Therefore, assuming $\inf \operatorname{dom} F > 0$, we can in (4.4) take $\gamma_N = \inf_{x \in \operatorname{dom} F} x\lambda$ for λ the infimum of the spectrum of the Laplacian as a bounded self-adjoint operator in $H^1(\Omega)$; see, e.g., [22, Theorem 9.2-1]. Clearly also $\alpha = 0$ due to $M_x = 0$. For (4.3), we get

$$\|A_x - A_{\tilde{x}}\|_{\mathbb{L}(U;W^*)} = \sup_{u,w}(x - \tilde{x})\frac{\langle \nabla u, \nabla w\rangle_{L^2(\Omega)}}{\|u\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}} = \sup_{u}(x - \tilde{x})\frac{\|\nabla u\|^2_{L^2(\Omega)}}{\|u\|_{H^1(\Omega)}}.$$

Thus we can take L_A as the supremum of the spectrum of the Laplacian as a bounded self-adjoint operator in $H^1(\Omega)$.

In the following examples, we take $U = W = \mathbb{R}^n$ with the standard Euclidean norm. Then (4.4) can be rewritten as the spectral radius bound and positivity condition

$$\rho(N_x^{-1}M_x), \rho(N_x^{-1,*}M_x^*) \le \alpha \quad \text{and} \quad N_x^*N_x \ge \gamma_N^2.$$

The first example also works in general spaces, as seen in a special case in Example 4.7, but γ_N and L_A depend on the norms chosen. Theorem 4.6 now shows that Assumption 3.3 holds.

Example 4.8 (No splitting). If $N_x = A_x \in \mathbb{R}^{n\times n}$, (4.4) holds with $\alpha = 0$ and γ_N the minimal eigenvalue of A_x, assumed symmetric positive definite. Theorem 4.6 now shows that Assumption 3.3 holds, where for any $\gamma_B > 1$ and $\beta > 0$, we can take $\pi_w = (1+\beta)\gamma_B\gamma_N^{-2}L_A^2\|\bar{w}\|^2$, $C_Q = (1+\beta^{-1})\gamma_B\gamma_N^{-2}L_Q^2$, and $\pi_u = \gamma_B\gamma_N^{-2}[(1+\beta)L_A^2\|\bar{u}\|^2 + (1+\beta^{-1})L_f^2]$.

Example 4.9 (Jacobi splitting). If N_x is the diagonal of $A_x \in \mathbb{R}^{n\times n}$, we obtain Jacobi splitting. The first part of (4.4) reduces to strict diagonal dominance, see [10, §10.1]. The second part always holds and N_x is invertible when the diagonal of A_x has only positive entries. Then γ_N is the minimum of the diagonal values. Theorem 4.6 now shows that Assumption 3.3 holds.

Example 4.10 (Gauss–Seidel splitting). If N_x is the lower triangle and diagonal of $A_x \in \mathbb{R}^{n\times n}$, we obtain Gauss–Seidel splitting. The first part of (4.4) holds for some $\alpha \in [0,1)$ when A_x is symmetric and positive definite; compare [10, proof of Theorem 10.1.2]. The second part holds for some γ_N when N_x is invertible. Theorem 4.6 now shows that Assumption 3.3 holds.

Example 4.11 (Successive over-relaxation). Based on any one of Examples 4.8 to 4.10, take $\tilde{N}_x = (1+r)N_x$ and $\tilde{M}_x = M_x - rN_x$ for some $r > 0$. Then, for small enough γ_B, all $\pi_u, \pi_w, C_Q \searrow 0$ as $r \nearrow \infty$.

Indeed, $\tilde{N}_x^{-1}\tilde{M}_x z = \tilde{\lambda} z$ if and only if $M_x z = ((1+r)\tilde{\lambda} + r)N_x z$, which gives the eigenvalues $\tilde{\lambda}$ of $\tilde{N}_x^{-1}\tilde{M}_x$ as $\tilde{\lambda} = (\lambda - r)/(1+r)$ for λ an eigenvalue of $N_x^{-1}M_x$. So, for large r, we can in (4.4) take $\alpha = (r+\rho)/(1+r)$ and $\gamma_{\tilde{N}} = \gamma_N(1+r)$, where $\rho := \rho(N_x^{-1}M_x) < 1$. Now, for every large enough $r > 0$, for $\gamma_B = (1+\alpha^{-2})/2 > 1$, we have

$$\begin{aligned}\frac{\alpha^2}{\gamma_{\tilde{N}_x}^2(1-\alpha^2\gamma_B)} = \frac{2\alpha^2}{\gamma_{\tilde{N}_x}^2(1-\alpha^2)} &= \frac{2(1+r)^2\alpha^2}{(1+r)^2\gamma_{\tilde{N}_x}^2((1+r)^2 - (1+r)^2\alpha^2)} \\ &= \frac{2(r+\rho)^2}{(1+r)^2\gamma_{\tilde{N}_x}^2((1+r)^2 - (r+\rho)^2)} = \frac{2(r+\rho)^2}{(1+r)^2\gamma_{\tilde{N}_x}^2(1-\rho^2+2(1-\rho)r)}.\end{aligned}$$

Since $0 \le \rho < 1$, the right hand side tends to zero as $r \nearrow \infty$. Since also $1/\gamma_N^2 \searrow 0$, and $\gamma_B > 1$, Theorem 4.6 now shows that Assumption 3.3 holds with $\pi_u, \pi_w, C_Q \searrow 0$ as $r \nearrow \infty$.

Quasi-conjugate gradients With $f_x = 0$ for simplicity, motivated by the conjugate gradient method for solving $A_x u = b$, see, e.g., [10], we propose to perform on Line 3 of Algorithm 2.1, and analogously

Line 4 the quasi-conjugate gradient update

$$\begin{cases} r^k := b - A_{x^k} u^k, \\ z^{k+1} := -\langle p^k, A_{x^k} r^k \rangle / \|p^k\|^2_{A_{x^k}}, \\ p^{k+1} := r^k + z^{k+1} p^k, \\ t^{k+1} := \langle p^{k+1}, r^k \rangle / \|p^{k+1}\|^2_{A_{x^k}}, \\ u^{k+1} := u^k + t^{k+1} p^{k+1}. \end{cases} \tag{4.6}$$

For standard conjugate gradients $A_{x^k} \equiv A$ permits a recursive residual update optimization that we are unable to perform. We have $\langle A_{x^k} p^{k+1}, p^k \rangle = 0$ for all k, although no "A-conjugacy" relationship necessarily exists between p^{k+1} and p^j for $j < k$.

The next lemma molds the updates (4.6) into our overall framework.

Lemma 4.12. *The update* (4.6) *corresponds to* Line 3 *of* Algorithm 2.1 *with*

$$\Gamma_k(u, \,\cdot\,, x) = \left[\mathrm{Id} - \|p^{k+1}\|^{-2}_{A_x} A_x \left(p^{k+1} \otimes p^{k+1} \right) \right] (A_x u^k - b) \quad (u \in U). \tag{4.7}$$

for $p^{k+1} = r^k_x + z^{k+1}_x p^k$ *for* $z^{k+1}_x = -\langle p^k, A_x r^k_x \rangle / \|p^k\|^2_{A_x}$ *and* $r^k_x := A_x u^k - b$.

Proof. Indeed, expanding t^{k+1}, the u-update of (4.6) may be rewritten as

$$u^{k+1} - u^k = \|p^{k+1}\|^{-2}_{A_{x^k}} (p^{k+1} \otimes p^{k+1}) r^k.$$

Applying the invertible matrix A_{x^k} and expanding r^k, this is

$$A_{x^k}(u^{k+1} - u^k) = -\|p^{k+1}\|^{-2}_{A_{x^k}} A_{x^k} (p^{k+1} \otimes p^{k+1})(A_{x^k} u^k - b),$$

and, adding $A_{x^k} u^k - b$ on both sides, further

$$A_{x^k} u^{k+1} - b = [\mathrm{Id} - \|p^{k+1}\|^{-2}_{x^k} A_{x^k} (p^{k+1} \otimes p^{k+1})](A_{x^k} u^k - b).$$

Since $B(u^{k+1}, \,\cdot\,; x^k) = \langle A_{x^k} u^{k+1}, \,\cdot\, \rangle$, and $L(\,\cdot\,) = \langle b, \,\cdot\, \rangle$, the claim follows. □

Unless A_x is independent of x, a simple approach as in Theorem 4.6 can only verify Assumption 3.3 with $\gamma_B < 1$. We hence leave the verification of convergence of Algorithm 2.1 with quasi-conjugate gradient updates to future research.

4.3 DISCUSSION

Before we embark on numerical experiments, it is time to make a few unifying observations about the disparate results above, with regard to the main conditions (3.16b) and (3.19b) of the convergence Theorems 3.10 and 3.11, and their connection to the fundamentally *discrete* viewpoint of Examples 4.9 and 4.10. As we have already noted in Remark 3.12,

(i) The main conditions (3.16b) and (3.19b) are easily satisfied for linear PDEs, i.e., when B_x does not depend on u. In Section 4.2, this corresponds to $A_x = A$ (while f_x may still depend on x). The only condition given in Remark 3.12 was that $\pi_w = 0$, which is satisfied in Examples 4.8 to 4.10 due to $L_A = 0$.

For linear PDEs, $\mathcal{S}(\bar w) = 0$. Together with $\pi_w = 0$, this causes also $\mathcal{S}(\bar u)$ and π_u to disappear from the convergence conditions. All of these quantities *might* depend on the discretisation.

As we have seen in Section 4.1, $\mathcal{S}(\bar u)$ and $\mathcal{S}(\bar w)$ require the use of ∞-norm bounds on the solutions, even when the underlying space is H^k. Such bounds may not always hold in infinite dimensions (however, see Remark 4.2), although they do always hold in finite-dimensional subspaces. In our numerical experiments, we have, however, not observed any grid dependency of $\mathcal{S}(\bar u)$ and $\mathcal{S}(\bar w)$ (calculated a posteriori, after a very large number of iterations).

On a more negative note, with $U = W = \mathbb{R}^{n(h,d)}$ equipped with the standard Euclidean norm, consider $A_x = -x\Delta_h$ for a scalar x with Δ_h a finite differences discretisation of the Laplacian on a d-dimensional square grid of cell width h and $n(h, d)$ nodes. Then, for both Jacobi and Gauss–Seidel splitting, as well as the trivial splitting (gradient descent) $N_x \propto \mathrm{Id}$, the spectral radius $\rho(N_x^{-1}M_x) \nearrow 1$ as $h \searrow 0$; see, e.g., [23, Chapter 4.2.1]. By simple numerical experiments, L_A^2/γ_N^2 nevertheless stays roughly constant, so the result is that $\pi_u, \pi_w \nearrow \infty$ as $h \searrow 0$. For “no splitting”, i.e., $N_x = A_x$, instead $L_A^2/\gamma_N^2 \nearrow \infty$ due to the worsening condition number of Δ_h. This latter negative result is, however, dependent on taking $U = W = \mathbb{R}^{n(h,d)}$ with the standard Euclidean norm: in Example 4.7 we showed that “no splitting” is applicable to the same problem in H^1. It is, therefore, an interesting question for future research, whether a change of norms would remove the grid dependency of Jacobi and Gauss–Seidel. Our guess is that it would not.

The above indicates that, for nonlinear PDEs, whether our methods even convergence, can depend on the level of discretisation. Nevertheless, to help comes the successive over-relaxation of Example 4.11, which shows that

(ii) By letting the over-relaxation parameter $r \nearrow \infty$, we get $\pi_u, \pi_w, C_Q \searrow 0$, and therefore may be able to obtain convergence (with a comparable iteration count) for any magnitude of $\mathcal{S}(\bar u)$, $\mathcal{S}(\bar w)$.

With over-relaxation $\gamma_B \searrow 1$ as $r \nearrow \infty$, so even then, to satisfy (3.16b) and (3.19b), it is necessary to have very small C_x. However,

(iii) In Sections 3 and 4.1, we have bounded C_x through $\operatorname{dom} F$, obtaining global convergence when (3.16b) and (3.19b) hold. With a more refined analysis, it is possible to make C_x arbitrary small by sufficiently good initialisation, i.e., by being content with mere local convergence.

We include a sketch of this analysis in Appendix B.

Finally, although convergence rates ($O(1/N^2)$ or linear) are unaffected by the discretisation level, constant factors of convergence depend on $Z_k\tilde{M}_k$ through the bound (3.17). This operator, written out in (3.13), depends on the constants π_u and π_w. They inversely scale the magnitude of the testing parameters λ_k and θ_k as chosen in (3.11). By (3.10), the term $\varphi_k + \lambda_k\pi_u + \theta_k\pi_w$ in (3.13) is, however, independent of π_u and π_w. Smaller π_u and π_w are, hence, better for the convergence of u and w (by weighing down the x and y initialisation errors on the right hand side of (3.17)), and higher π_u and π_w are better for the convergence of x and y (by weighing down u and w initialisation errors). Even for linear PDEs, therefore

(iv) Convergence speed may depend on the level of discretisation through the x-sensitivity factors π_u and π_w of the splitting method for the PDE.

This is to be expected: the linear system solvers that Section 4.2 is based on, are *fundamentally discrete*, and their convergence depends on the eigenvalues of $N_x^{-1}M_x$ and N_x. In “standard” optimisation methods, the dimensionally-dependent linear system solver is taken as a black box, and its computational cost is hidden from the estimates for the optimisation method. The estimates for our method, by contrast, include the solver.

5 NUMERICAL RESULTS

We now illustrate the numerical performance of Algorithm 2.1. We first describe our experimental setup, and then discuss the results.

5.1 EXPERIMENTAL SETUP

The PDEs in our numerical experiments take one of the forms of Section 4.1 on the domain $\Omega = [0,1] \times [0,1]$ with nonhomogeneous Dirichlet boundary conditions. We discretize the domain as a regular grid and the PDEs by backward differences. We use both a coarse and a fine grid.

The function G and the PDE vary by experiment, but in each one we take the regularization term for the control parameter x and the data fitting term as

$$(5.1) \qquad F(x) := \frac{\alpha}{2}\|x\|^2_{L^2(\Omega;\mathbb{R}^{d\times d})\times L^2(\Omega)} + \delta_{[\lambda,\lambda^{-1}]}(x) \quad \text{and} \quad Q(u) := \widehat{\beta}\sum_{i=1}^m \|u_i - z_i\|^2_{L^2(\Omega)}$$

for some $\alpha, \beta, \lambda > 0$ as well as $\widehat{\beta} := \beta/(2\|\bar{z}\|^2_{L^2(\Omega)})$ where $\bar{z} = \frac{1}{m}\sum_{i=1}^m z_i$ is the average of the measurement data z_i. The norms here are in function spaces, but in the numerical experiments the variables are, of course, taken to be in a finite-dimensional (finite element) subspace.

The variables u_i correspond to multiple copies of the same PDE with different boundary conditions $u_i = f_i$ on $\partial\Omega$, $(i = 1, \ldots, m)$, for the same control x. Parametrizing $\partial\Omega$ by $\rho : (0,1) \to \partial\Omega$, we take as boundary data

$$(5.2) \qquad f_{2j-1}(\rho(t)) = \cos(2\pi jt) \quad \text{and} \quad f_{2j}(\rho(t)) = \sin(2\pi jt), \quad (j = 1, \ldots, m/2).$$

To produce the synthetic measurement z_i, we solve for $\hat{u}_i$ the PDE corresponding to the experiment with the ground truth control parameter $\hat{x} = (\hat{A}, \hat{c})$ and boundary data f_i. To this we add Gaussian noise of standard deviation $0.01\|\hat{u}_i\|_{L^2(\Omega)}$ to get z_i.

We next describe the PDEs for each of our experiments.

Experiment 1 (Scalar coefficient). In our first numerical experiment, we aim to determine the scalar coefficient $c \in \mathbb{R}$ for the PDEs

$$(5.3) \qquad \begin{cases} -\Delta u_i + cu_i = 0 & \text{in } \Omega, \\ u_i = f_i & \text{on } \partial\Omega, \end{cases}$$

where $i = 1, \ldots, m$. For this problem we choose $G(Kx) = 0$. Thus the objective is

$$(5.4) \qquad \min_{u,c} J(x) := \frac{\alpha}{2}\|c\mathbf{1}\|^2_{L^2(\Omega)} + \delta_{[\lambda,\lambda^{-1}]}(c) + \widehat{\beta}\sum_{i=1}^m \|u_i - z_i\|^2_{L^2(\Omega)} \quad \text{subject to (5.3).}$$

Our parameter choices can be found in Table 1.

With $u = (u_1, \ldots, u_m) \in U^m \subset H^1(\Omega)^m$ and $w = (w_{1,\Omega}, \ldots, w_{m,\Omega}, w_{1,\partial}, \ldots, w_{m,\partial}) \in W^m \subset H^1_0(\Omega)^m \times H^{1/2}(\partial\Omega)^m$, for the weak formulation of (5.3) we take

$$B(u, w; c) = \sum_{i=1}^m \left(\langle \nabla u_i, \nabla w_{i,\Omega}\rangle_{L^2(\Omega)} + c\langle u_i, w_{i,\Omega}\rangle_{L^2(\Omega)} + \langle \operatorname{trace}_{\partial\Omega} u_i, w_{i,\partial}\rangle_{L^2(\partial\Omega)}\right)$$

and

$$(5.5) \qquad Lw = \sum_{i=1}^m \langle f_i, w_{i,\partial}\rangle_{L^2(\partial\Omega)}.$$

Then $\bar{\nabla}_x B(u,w) = \sum_{i=1}^m \langle u_i, w_{i,\Omega}\rangle_{L^2(\Omega)}$ following Example 4.5.

For data generation we take $\hat{c} = 1.0$. Since we are dealing with an ill-posed inverse problem, an optimal control parameter $\bar{c}$ for (5.4) does not in general equal $\hat{c}$. Therefore, to compare algorithm progress, we take as surrogate for the unknown $\bar{c}$ the iterate $\tilde{c}_A := c^{50{,}000}$ on the coarse grid and $\tilde{c}_B := c^{500{,}000}$ on the fine grid, each computed using Algorithm 2.1 without splitting.

The next theorem verifies the basic structural conditions of the convergence Theorems 3.10 and 3.11. The splitting conditions contained Assumption 3.3 are ensured through Example 4.9 (Jacobi), 4.10 (Gauss–Seidel), or 4.8 (no splitting).

Theorem 5.1. *Let $X = \mathbb{R}$; U a finite-dimensional subspace of $H^1(\Omega)$; and W a finite-dimensional subspace of $H_0^1(\Omega) \times H^{1/2}(\partial\Omega)$. Let F and Q be given by* (5.1) *along with the PDE* (5.3) *and the boundary conditions f_i defined as in* (5.2)*. Take $G = 0$. Then Assumption 3.1 holds.*

Proof. The chosen F, Q and either G satisfy Assumption 3.1(i). The boundary conditions $f_i \in H^{1/2}(\partial\Omega)$ along with the constraint $x \in [\lambda, \lambda^{-1}]$ ensure the condition Lemma 4.1(ii′). In the discretized setting, also (iii′) and (iv′) also hold. In conclusion, Lemma 4.1 verifies Assumption 3.1. □

Remark 5.2. It remains to verify (3.16) or (3.19), depending on the convergence theorem used. The condition (3.16a) is readily verified by appropriate choice of the primal and dual step length parameters $\tau_0, \sigma_0 > 0$. We also take $\tilde{\gamma}_F = 0$ (slightly violating the assumptions), so that $\omega_k \equiv 1$, and $\tau_k \equiv \tau_0$ and $\sigma_k \equiv \sigma_0$. The condition (3.16b) (and likewise (3.19b) for linear convergence) is very difficult to verify *a priori* for nonlinear PDEs, as it depends on the knowledge of a solution to the optimisation problem through $\mathcal{S}(\bar{u})$ and $\mathcal{S}(\bar{w})$. This is akin to the difficulty of verifying (a priori) a positive Hessian at a solution for standard nonconvex optimisation methods. Hence we do not attempt to verify (3.16b).

Experiment 2 (Diffusion + scalar coefficient). In this experiment we aim to determine the coefficient function $a : \Omega \to \mathbb{R}$ and scalar $c \in \mathbb{R}$ for the group of PDEs

$$\begin{cases} -\nabla \cdot (a\nabla u_i) + cu_i = 0 & \text{in } \Omega, \\ u_i = f_i & \text{on } \partial\Omega, \end{cases} \tag{5.6}$$

where $i = 1, \ldots, m$. The optimization problem then is

$$\min_{x=(a,c)} J(x) = \delta_{[\lambda,\lambda^{-1}]}(x) + \widehat{\beta} \sum_{i=1}^m \|u_i - z_i\|^2_{L^2(\Omega)} + \gamma \|\nabla a\|_1 \quad \text{subject to (5.6)}. \tag{5.7}$$

Due to the total variation term, we need to ensure that a is in H^1 through discretisation. Moreover, for any control (a, c), corresponding solutions u_i to (5.6) solves the PDEs for (ta, tc) for any $t > 0$. In our progress plots we therefore consider $\frac{a}{c}$.

For the weak formulation of (5.6) with $w = (w_{1,\Omega}, \ldots, w_{m,\Omega}, w_{1,\partial}, \ldots, w_{m,\partial}) \in W^m \subset H_0^1(\Omega)^m \times H^{1/2}(\partial\Omega)^m$, $u = (u_1, \ldots, u_m) \in U^m \subset H^1(\Omega)^m$, and $x = (a, c) \in X \subset L^2(\Omega) \times \mathbb{R}$, we take L as in (5.5) and

$$B(u, w; x) = \sum_{i=1}^m \left(\langle \nabla u_i, a\nabla w_{i,\Omega}\rangle_{L^2(\Omega)} + c\langle u_i, w_{i,\Omega}\rangle_{L^2(\Omega)} + \langle \operatorname{trace}_{\partial\Omega} u_i, w_{i,\partial}\rangle_{L^2(\partial\Omega)} \right).$$

Then $\bar{\nabla}_x B(u,w) = (\bar{\nabla}_x B^1(w,u), \bar{\nabla}_x B^2(w,u))$ takes on a mixed form with $\bar{\nabla}_x B^1(w,u) = \sum_{i=1}^m \nabla u_i \cdot \nabla w_{i,\Omega}$ from Example 4.4 and $\bar{\nabla}_x B^2(w,u) = \sum_{i=1}^m \langle u_i, w_{i,\Omega}\rangle_{L^2(\Omega)}$ from Example 4.5.

For data generation we take $\hat{c} = 1.0$ and $\hat{a}$ as the phantom in Figure 5. Similarly to Experiment 1 we compare the progress towards $\tilde{a} := a^{1{,}000{,}000}$ and $\tilde{c} := c^{1{,}000{,}000}$ computed using Algorithm 2.1 with full matrix inversion.

As above for Experiment 1, the next theorem verifies the basic structural conditions of the convergence Theorems 3.10 and 3.11. The proofs is analogous to that Theorem 5.1. Likewise, the splitting

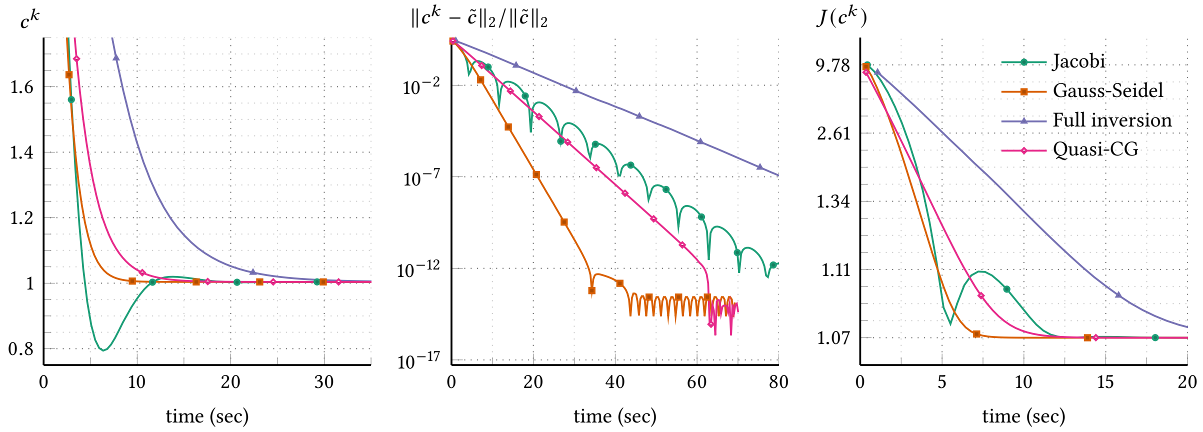

Figure 1: Performance of various splittings in the coarse grid Experiment 1.

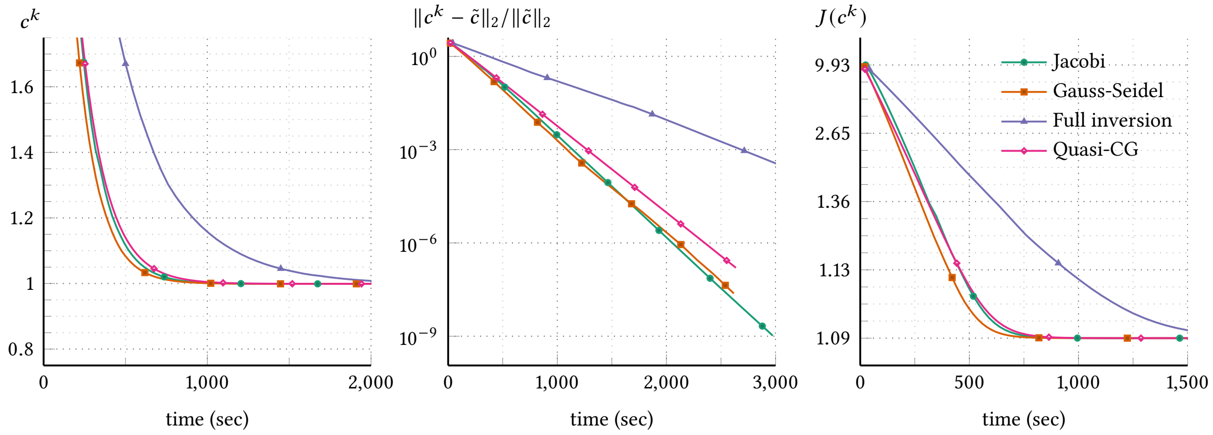

Figure 2: Performance of various splittings in fine grid Experiment 1.

Assumption 3.3 is verified as before through Example 4.9 (Jacobi), 4.10 (Gauss–Seidel), or 4.8 (no splitting), while Remark 5.2 applies for the remaining step length and growth conditions.

Theorem 5.3. *Let X be a finite-dimensional subspace of $L^2(\Omega) \times \mathbb{R}$, U a finite-dimensional subspace of $H^1(\Omega)$ and W a finite-dimensional subspace of $H_0^1(\Omega) \times H^{1/2}(\partial\Omega)$. Let F and Q be given by (5.1) along with the PDE (5.6) with the boundary conditions f_i defined as in (5.2) and G be $\|\cdot\|_1$. Then Assumption 3.1 holds.*

5.2 ALGORITHM PARAMETRISATION

We apply Algorithm 2.1 with no splitting (full inversion), and with Jacobi and Gauss–Seidel splitting, and quasi conjugate gradients, as discussed in Section 4.2. We fix $\sigma = 1.0$, $\omega = 1.0$, $\lambda = 0.1$, $\varepsilon = 0.01$, and $\beta = 10^2$ for all experiments. Other parameters, including the grid size, α, γ_i, τ and m vary according to experiment with values listed in Table 1.

For the initial iterate (x^0, u^0, w^0, y^0) we make an experiment-specific choice of the control parameter x^0. Then we determine u^0 by solving the PDE, and w_0 by solving the adjoint PDE. We set $y^0 = Kx^0$. For Experiment 1 we take the initial $c^0 = 4.0$ and run the algorithm for 20,000 iterations on the coarse grid and 125,000 on the fine. For Experiment 2 we take the initial $a^0 \equiv 1.0$ a constant function, and $c^0 = 2.0$. The algorithm is run for 200,000 iterations on the coarse grid, and 500,000 on the fine.

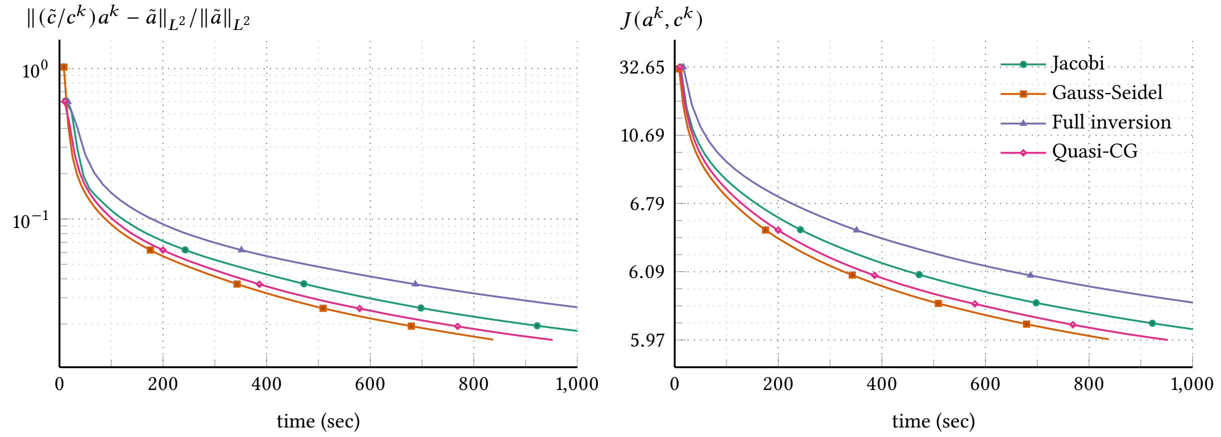

Figure 3: Performance of various splittings in the coarse grid Experiment 2.

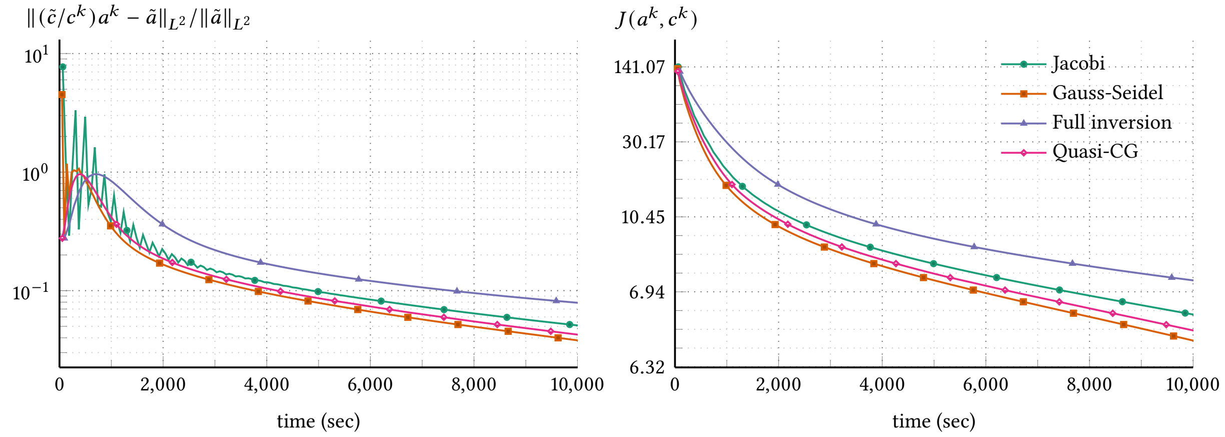

Figure 4: Performance of various splittings in the fine grid Experiment 2.

We implemented the algorithm in Julia. Our implementation is available at doi:10.5281/zenodo.7389545 on Zenodo. The experiments were run on a ThinkPad laptop with Intel Core i5-8265U CPU at 1.60GHz ×4 and 15.3 GiB memory.

5.3 RESULTS

The results for Experiment 1 with the above algorithm parametrisations are in Figure 1 for the coarse grid and Figure 2 for the fin grid. In the figures we illustrate the evolution of the coefficient c^k as the algorithm iterates. We also show the evolution of the relative error of the coefficient and the functional value.

Table 1: Parameter choices for all examples.

Grid	N	Grid size	α	β	γ	τ	σ	ω	m
Coarse	51	2601	1×10^{-5}	1×10^{2}	0	2.5×10^{-2}	1	1	6
Fine	101	10201	1×10^{-5}	1×10^{2}	0	2.0×10^{-3}	1	1	6
Coarse	51	2601	0	1×10^{2}	10^{-2}	2.5×10^{-2}	1	1	10
Fine	101	10201	0	1×10^{2}	10^{-2}	1×10^{-2}	1	1	10

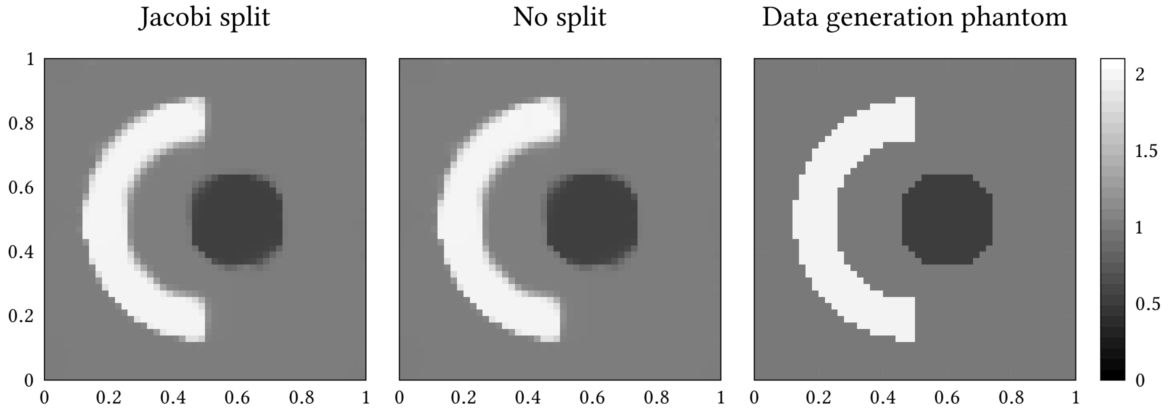

Figure 5: Illustrations of the coefficient reconstructions for Experiment 2A. On the left is the result of the Jacobi split approach, in the middle the full matrix inversion after the same number of iterations. On the right we show the data generation phantom for comparison.

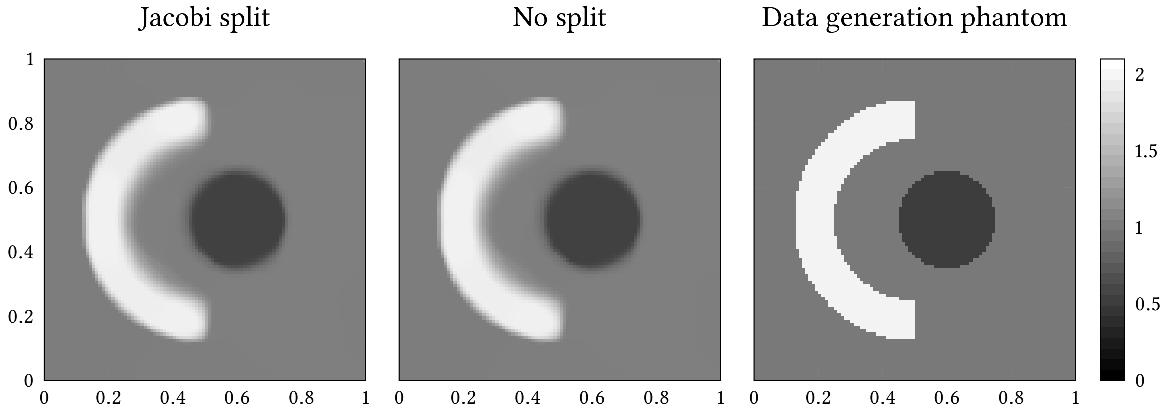

Figure 6: Illustrations of the coefficient reconstructions for Experiment 2B. On the left is the result of the Jacobi split approach, in the middle the full matrix inversion after the same number of iterations. On the right we show the data generation phantom for comparison.

The results for Experiment 2 are available in Figures 3 and 5 for the coarse grid and Figures 4 and 6 for the find grid. In Figures 3 and 4 are shown the evolution of the relative error of the coefficient and the functional value. In Figures 5 and 6 are the reconstructed coefficients a^k at the final iterates and for comparison the phantom used for the data generation.

The performance plots have *time* on the x-axis rather than the number of iterations, as the main difference between the splittings is expected to be in the computational effort for linear system solution, i.e., Lines 3 and 4 of Algorithm 2.1. For fairness, we limited the number of threads used by Julia/OpenBLAS to one.

In all experiments the splittings outperform full matrix inversion: the best splittings require roughly *half of the computational effort* for an iterate of the same quality. No particular splitting completely dominates another, however, Jacobi appear to be more prone to overstepping and oscillatory patterns. On the other hand, quasi-CG currently has no convergence theory, and we have observed situations where it does not exhibit convergence while Jacobi and Gauss–Seidel splittings do. Therefore, Gauss–Seidel is our recommended option.

APPENDIX A OPTIMALITY CONDITIONS

We prove here the necessity of (2.9) for solutions to (2.8).

Proof of Theorem 2.3. We let $T(x,u) := B(u,\,\boldsymbol{\cdot}\,;x)$, $T : X \times U \to W^*$. Setting

$$A := \{(x,u) \in X \times U \mid B(u,w;x) = Lw \text{ for all } w \in W\} = T^{-1}(L),$$

any solution $(\bar{u}, w, \bar{x}, \bar{y})$ to (2.8) also solves

$$\min_{x,u} R(x,u) := [R_0 + \delta_A](x,u) \quad \text{where} \quad R_0(x,u) = F(x) + Q(u) + G(Kx).$$

with $G(K\bar{x}) = \langle K\bar{x}, \bar{y}\rangle_Y - G^*(\bar{y})$. By the Fenchel-Young theorem, the latter is equivalent to the last line of (2.9). Clearly $(\bar{x},\bar{u}) \in A$, or else there is no solution. Therefore also the first line of (2.9) holds.

It follows from the linearity/affinity and continuity, hence continuous differentiability of B that T is strictly differentiable. Since $T'(\bar{x},\bar{u})(h_x,h_u) = B_x(\bar{u},\,\boldsymbol{\cdot}\,;h_x) + B_u(h_u,\,\boldsymbol{\cdot}\,;\bar{x})$, so that

$$\langle T'(\bar{x},\bar{u})^* w | (h_x,h_u)\rangle = B_x(\bar{u},w;h_x) + B_u(h_u,w;\bar{x}),$$

the qualification condition (2.10a) reads

$$\sup_{\|(h_x,h_u)\|=1} \|T'(\bar{x},\bar{u})^*(h_x,h_u)\| \geq c\|w\| \quad \text{for all} \quad w \in W.$$

Moreover, as a bounded linear operator, $T'(\bar{x},\bar{u})$ is closed, i.e., has closed graph. Therefore, by [2, Theorem 2.20], $T'(\bar{x},\bar{u})$ is surjective. With this, [26, Theorem 1.17] gives

$$\begin{aligned}\partial_M \delta_A(x,u) &= T'(\bar{x},\bar{u})^* N_{\{L\}}(T(\bar{x},\bar{u})),\\ &= \{(h_x,h_u) \mapsto \langle T'(\bar{x},\bar{u})(h_x,h_u)|w\rangle \mid w \in W\}\\ &= \{(h_x,h_u) \mapsto B_x(\bar{u},w;h_x) + B_u(h_u,w;\bar{x}) \mid w \in W\}.\end{aligned}$$

Here we denote by $N_D(x) = \partial_M \delta_D(x)$ the limiting normal cone to a set D at x.

Since limiting subdifferentials agree with convex subdifferentials on convex functions, and we have assumed that $\operatorname{int}\operatorname{dom} R_0 \neq \emptyset$, we can easily calculate $\partial_M R_0$. We will then use the sum rule [26, Theorem 3.36] to estimate $\partial_M R$, which requires verifying that R_0 is “sequentially normally epicompact” (SNEC), and that the "horizon subdifferentials", defined for $V : X \to \overline{\mathbb{R}}$ as $\partial^\infty V(x) := \{x^* \in X^* \mid (x^*, 0) \in N_{\operatorname{epi} V}(x, V(x))\}$, satisfy

$$\partial^\infty \delta_A(\bar x, \bar u) \cap (-\partial^\infty R(\bar x, \bar u)) = \{0\}. \tag{A.1}$$

Indeed, convex functions whose domains have a non-empty interior, such as R_0, are SNEC by [26, Proposition 1.25 and discussion after Definition 1.116]. Moreover, since $\partial^\infty Q(\bar u) = \{0\}$, (A.1) reduces to

$$B_u(\,\cdot\,, w; \bar x) = 0 \implies B_x(\bar u, w; \,\cdot\,) \cap (-\partial^\infty [F + G \circ K](x)) = \{0\}$$

This is guaranteed by the qualification condition (2.10b). Now, by the Fermat principle [26, Proposition 1.114] and the sum rule [26, Theorem 3.36], we have

$$0 \in \partial_M R(\bar x, \bar u) \subset \begin{pmatrix} \partial F(\bar x) + K^* \partial G(K \bar x) \\ \{Q'(\bar u)\} \end{pmatrix} + \partial_M \delta_A(\bar x, \bar u).$$

After appropriate Riesz representations, this inclusion expands as the middle two lines of (2.9). □

APPENDIX B LOCALIZATION

Theorems 3.10 and 3.11 are global convergence results, but also depend on the global constant C_x in Assumption 3.1 (iv). To satisfy the conditions of the theorems, $\operatorname{dom} F$ may need to be small for C_x to be small. We now develop local convergence results that allow replacing C_x by a small initialization-dependent value without restricting $\operatorname{dom} F$.

We replace Assumption 3.1 with the following:

Assumption B.1. We assume Assumption 3.1 to hold with (iv) replaced by

(iv′) For some $\tilde C_x \geq 0$, for all $(u, w) \in U \times W$ and $x \in \operatorname{dom} F$ we have the bound

$$B_x(u, w; x - \bar x) \leq \tilde C_x \|x - \bar x\|_X \|u\|_U \|w\|_W.$$

This estimate uses the standard norm in X, which is a 2-norm in the examples of Sections 4.1 and 5. However, Section 4.1 gives estimates involving an ∞-norm for C_x. Therefore some finite-dimensionality of the parameters is required to satisfy Assumption B.1 (iv′). This can take the form of a finite element discretisation of a function parameter a, or the parameter being a scalar constant. In the latter case, the examples of Section 4.1 readily verify Assumption B.1.

We then modify several previous results accordingly:

Lemma B.2 (Local version of Lemma 3.6). *Let $k \in \mathbb{N}$. Suppose Assumptions 3.3, 3.4 and B.1 hold,*

$$\|u^{k+1} - \bar u\|_U \leq \delta_{uw}, \quad \textit{and} \quad \|w^{k+1} - \bar w\|_U \leq \delta_{uw}, \tag{B.1}$$

for some $\delta_{uw} > 0$, and for some $\varepsilon_u, \varepsilon_w, \mu > 0$ that

$$\gamma_F \geq \tilde\gamma_F + \varepsilon_u + \varepsilon_w + \frac{\lambda_{k+1}\pi_u + \theta_{k+1}\pi_w}{\eta_k} \tag{B.2a}$$

$$\gamma_{G^*} \geq \tilde\gamma_{G^*}, \tag{B.2b}$$

$$\gamma_B \geq \frac{\lambda_{k+1}}{\lambda_k} + \frac{\theta_k}{\lambda_k} C_Q + \frac{\eta_k \mathcal{S}(\bar w)}{2\varepsilon_w \lambda_k} + \frac{\tilde C_x^2 \delta_{uw}^2 \mu \eta_k}{4\varepsilon_u \lambda_k}, \quad \textit{and} \tag{B.2c}$$

$$\gamma_B \geq \frac{\theta_{k+1}}{\theta_k} + \frac{\eta_k \mathcal{S}(\bar u)}{2\varepsilon_u \theta_k} + \frac{\tilde C_x^2 \delta_{uw}^2 \eta_k}{4\varepsilon_w \mu \theta_k}. \tag{B.2d}$$

Then (3.6) holds.

Proof. We follow the proof of Lemma 3.6 until the estimate (3.8), which now holds with $C_x = \tilde{C}_x\|x - \bar{x}\|_X$ and any $\tilde{\varepsilon}_u, \tilde{\varepsilon}_w, \tilde{\mu} > 0$ standing for $\varepsilon_u, \varepsilon_w, \mu > 0$. Recall that we abbreviate $u = u^{k+1}$, $w = w^{k+1}$, and $x = x^{k+1}$. Using Young's inequality and (B.1), we continue from there estimating that

$$\begin{aligned}
&\eta_k\langle\bar{\nabla}_x B(u, w) - \bar{\nabla}_x B(\bar{u}, \bar{w}), x - \bar{x}\rangle \\
&\quad\geq -\eta_k\left(\frac{\mathcal{S}(\bar{u})}{4\tilde{\varepsilon}_u} + \frac{\tilde{C}_x\|x - \bar{x}\|\tilde{\mu}}{2}\right)\|w - \bar{w}\|_W^2 - \eta_k\left(\frac{\mathcal{S}(\bar{w})}{4\tilde{\varepsilon}_w} + \frac{\tilde{C}_x\|x - \bar{x}\|}{2\tilde{\mu}}\right)\|u - \bar{u}\|_U^2 \\
&\qquad - \eta_k(\tilde{\varepsilon}_u + \tilde{\varepsilon}_w)\|x - \bar{x}\|_X^2 \\
&\quad\geq -\eta_k\left(\frac{\mathcal{S}(\bar{u})}{4\tilde{\varepsilon}_u} + \frac{\tilde{C}_x^2\delta_{uw}^2\tilde{\mu}^2}{8\tilde{\varepsilon}_u}\right)\|w - \bar{w}\|_W^2 - \eta_k\left(\frac{\mathcal{S}(\bar{w})}{4\tilde{\varepsilon}_w} + \frac{\tilde{C}_x^2\delta_{uw}^2}{8\tilde{\mu}^2\tilde{\varepsilon}_w}\right)\|u - \bar{u}\|_U^2 \\
&\qquad - \eta_k(2\tilde{\varepsilon}_u + 2\tilde{\varepsilon}_w)\|x - \bar{x}\|_X^2.
\end{aligned}$$

With $\varepsilon_u = 2\tilde{\varepsilon}_u$, $\varepsilon_u = 2\tilde{\varepsilon}_w$, and $\mu = \tilde{\mu}^2$, we now continue with the proof of Lemma 3.9, which goes through with (B.2) in place of (3.5). □

Lemma B.3 (Local version of Lemma 3.7). *Suppose $\gamma_F > \tilde{\gamma}_F > 0$ as well as $\gamma_{G^*} \geq \tilde{\gamma}_{G^*} \geq 0$ and that there exist $\omega, t > 0$ with $\omega\eta_{k+1} \leq \eta_k$ for all $k \in \mathbb{N}$ such that*

$$\gamma_B \geq \omega^{-1} + tC_Q + \frac{4(1 + t^{-1})}{\omega(\gamma_F - \tilde{\gamma}_F)^2}\left(\mathcal{S}(\bar{u})\pi_w + t\mathcal{S}(\bar{w})\pi_u + \frac{1}{4}\sqrt{t\pi_w\pi_u}\tilde{C}_x^2(\gamma_F - \tilde{\gamma}_F)\delta_{uw}^2\right). \tag{B.3}$$

Then there exist $\varepsilon_u, \varepsilon_w, \mu > 0$, and, for all $k \in \mathbb{N}$, $\lambda_k, \theta_k > 0$, such that (B.2) holds.

Proof. In the proof of Lemma 3.7, we replace C_x by $\tilde{C}_x^2\delta_{uw}^2$, and use (B.2) in place of (3.5) and (B.3) in place of (3.9). Observe that compared to (3.5c) and (3.5d), (B.2c) and (B.2d) have an additional factor 2 in front of the terms involving ε_u and ε_w. This difference produces the constant factors 4 instead of 2 in (B.3) compared to (3.9). □

Lemma B.4 (Local version of Lemma 3.9). *Suppose Assumptions 3.4 and B.1 hold as do Assumption 3.3 and (B.2) for $k = 0, \ldots, N - 1$ with*

$$\delta_{uw}^2 = \frac{1}{\gamma_B}\max\left\{\frac{1}{\lambda_0}, \frac{C_Q\gamma_B^{-1}}{\lambda_0}, \frac{1}{\theta_0}, \frac{1 + C_Q\gamma_B^{-1}}{\lambda_0 + \theta_0}\right\}\delta^2 \tag{B.4}$$

and

$$\delta := \|v^0 - \bar{v}\|_{Z_0\tilde{M}_0}. \tag{B.5}$$

Also suppose $\{\lambda_k\}_{k\in\mathbb{N}}$ and $\{\theta_k\}_{k\in\mathbb{N}}$ are non-decreasing. Given v^0, let $v^1, \ldots, v^{N-1}$ be produced by Algorithm 2.1. Then (3.14) holds for $k = 0, \ldots, N - 1$, where all the terms are non-negative.

Proof. We need to prove (B.1) for all $k = 0, \ldots, N - 1$. The rest follows as in the proof of Lemma 3.9.

Assumption 3.3 (iii) with (3.13) and Lemma 3.5 establish for all $k = 0, \ldots, N - 1$ the *a priori* bounds

$$\begin{aligned}
\|u^{k+1} - \bar{u}\|_U^2 &\leq \frac{1}{\gamma_B}\left(\|u^k - \bar{u}\|_U^2 + \pi_u\|x^k - \bar{x}\|_X^2\right) \\
&\leq \frac{1}{\gamma_B}\max\left\{\frac{1}{\lambda_k}, \frac{\pi_u}{\varphi_k(1 - \kappa) + (\lambda_k + \theta_k)\pi_u}\right\}\|v^k - \bar{v}\|_{Z_k\tilde{M}_k}^2 \\
&\leq \frac{1}{\gamma_B}\max\left\{\frac{1}{\lambda_0}, \frac{1}{\lambda_0 + \theta_0}\right\}\|v^k - \bar{v}\|_{Z_k\tilde{M}_k}^2 \\
&\leq \frac{\delta_{uw}^2}{\delta^2}\|v^k - \bar{v}\|_{Z_k\tilde{M}_k}^2
\end{aligned} \tag{B.6}$$

and

$$
\begin{aligned}
\|w^{k+1}-\bar w\|_W^2 &\le \frac{1}{\gamma_B}\left(\|w^k-\bar w\|_W^2 + C_Q\|u^{k+1}-\bar u\|_U^2 + \pi_w\|x^k-\bar x\|_X^2\right)\\
&\le \frac{1}{\gamma_B}\left(\|w^k-\bar w\|_W^2 + C_Q\gamma_B^{-1}\|u^k-\bar u\|_U^2 + (1+C_Q\gamma_B^{-1})\pi_w\|x^k-\bar x\|_X^2\right)\\
&\le \frac{1}{\gamma_B}\max\left\{\frac{1}{\theta_k}, \frac{C_Q\gamma_B^{-1}}{\lambda_k}, \frac{(1+C_Q\gamma_B^{-1})\pi_w}{\varphi_k(1-\kappa)+(\lambda_k+\theta_k)\pi_w}\right\}\|v^k-\bar v\|_{Z_k\tilde M_k}^2\\
&\le \frac{1}{\gamma_B}\max\left\{\frac{1}{\theta_0}, \frac{C_Q\gamma_B^{-1}}{\lambda_0}, \frac{1+C_Q\gamma_B^{-1})}{\lambda_0+\theta_0}\right\}\|v^k-\bar v\|_{Z_k\tilde M_k}^2\\
&\le \frac{\delta_{uw}^2}{\delta^2}\|v^k-\bar v\|_{Z_k\tilde M_k}^2.
\end{aligned}
\tag{B.7}
$$

In the final steps we have used the the assumptions that $\{\varphi_k\}_{k\in\mathbb N}$ (by Assumption 3.4), $\{\lambda_k\}_{k\in\mathbb N}$, and $\{\theta_k\}_{k\in\mathbb N}$ are non-decreasing.

We now use induction. By definition we have $\|v^0-\bar v\|_{Z_0\tilde M_0}\le\delta$. Hence (B.6) and (B.7) verify (B.1) for $k=0$. Suppose then that we have proved (B.1) for $k=0,\ldots,\ell-1$. Then (3.14) holds $k=0,\ldots,\ell-1$ by following the proof of Lemma 3.9, replacing Lemma 3.6 there in by the localized Lemma B.2. Summing (3.14) over $k=0,\ldots,\ell-1$, we now obtain the *a posteriori* bound

$$
\frac{1}{2}\|v^\ell-\bar v\|_{Z_\ell\tilde M_\ell}^2 \le \frac{1}{2}\|v^0-\bar v\|_{Z_0\tilde M_0}^2 = \frac{1}{2}\delta^2.
$$

Now (B.6) and (B.7) verify (B.1) for $k=\ell$. Hence also (3.14) holds for $k=\ell$. As a result of the entire inductive argument, it holds for all $k=0\ldots,N-1$. □

With $\varphi_0=1$ and the choices of

$$
\lambda_0 := t^{-1}r_0\pi_u^{-1}\eta_0 \quad\text{and}\quad \theta_0 := r_0\pi_w^{-1}\eta_0 \quad\text{for}\quad r_0 := \frac{\gamma_F-\tilde\gamma_F}{2(t^{-1}+1)c_0}
$$

and $c_0 := \eta_1/\eta_0 \le \omega^{-1}$ in the proof of Lemma B.3 (Lemma 3.7), we expand and estimate (B.4) as

$$
\begin{aligned}
\delta_{uw}^2 &= \frac{1}{\gamma_B}\max\left\{\frac{t\pi_u}{r_0\eta_0}, \frac{tC_Q\pi_u}{r_0\eta_0\gamma_B}, \frac{\pi_w}{r_0\eta_0}, \frac{1+C_Q\gamma_B^{-1}}{(t^{-1}\pi_u^{-1}+\pi_w^{-1})r_0\eta_0}\right\}\delta^2\\
&= \frac{t}{\gamma_B r_0\eta_0}\max\left\{\pi_u, \frac{C_Q\pi_u}{\gamma_B}, \frac{\pi_w}{t}, (1+C_Q\gamma_B^{-1})\frac{\pi_u\pi_w}{\pi_w+t\pi_u}\right\}\delta^2\\
&\le \frac{2(1+t)}{\gamma_B\eta_0\omega(\gamma_F-\tilde\gamma_F)}\max\left\{\pi_u, \frac{C_Q\pi_u}{\gamma_B}, \frac{\pi_w}{t}, (1+C_Q\gamma_B^{-1})\pi_u\right\}\delta^2.
\end{aligned}
\tag{B.8}
$$

Hence (B.3) with δ_{uw}^2 replaced by this upper estimate and $\varphi_0=1$ (so that $\eta_0=\tau_0$) reads

$$
\gamma_B \ge \omega^{-1} + tC_Q + \frac{4(1+t^{-1})}{\omega(\gamma_F-\tilde\gamma_F)^2}\left(\mathcal S(\bar u)\pi_w + t\mathcal S(\bar w)\pi_u + \frac{\max\left\{\pi_u, \frac{C_Q\pi_u}{\gamma_B}, \frac{\pi_w}{t}, (1+C_Q\gamma_B^{-1})\pi_u\right\}(1+t)\sqrt{t\pi_w\pi_u}}{2\gamma_B\tau_0\omega}\tilde C_x^2\delta^2\right).
\tag{B.9a}
$$

where we recall that $t>0$ is a free balancing parameter, and

$$
\delta := \|v^0-\bar v\|_{Z_0\tilde M_0}.
\tag{B.9b}
$$

We now immediately obtain local versions of the main results. By initializing close enough to a solution, i.e., with small δ, we can possibly obtain convergence more often than from the global versions.

Corollary B.5 (Local accelerated convergence). *In Theorem 3.10, replace Assumption 3.1 by Assumption B.1 and* (3.16b) *by* (B.9) *with $\omega = \omega_0$. Then the claims continue to hold.*

Corollary B.6 (Local linear convergence). *In Theorem 3.11, replace Assumption 3.1 by Assumption B.1 and* (3.19b) *and* (B.9) *with $\tau_0 = \tau$. Then the claims continue to hold.*

Both proofs are exactly as the original proofs, using Lemma B.4 in place of Lemma 3.9.

REFERENCES

[1] T. Bosse, N. R. Gauger, A. Griewank, S. Günther, and V. Schulz, One-Shot Approaches to Design Optimization, *Trends in PDE Constrained Optimization* (2014), 43–66, doi:10.1007/978-3-319-05083-6_5.

[2] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2011, doi:10.1007/978-0-387-70914-7.

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

[4] C. Clason, S. Mazurenko, and T. Valkonen, Acceleration and global convergence of a first-order primal-dual method for nonconvex problems, *SIAM Journal on Optimization* 29 (2019), 933–963, doi:10.1137/18m1170194, arXiv:1802.03347.

[5] C. Clason, S. Mazurenko, and T. Valkonen, Primal-dual proximal splitting and generalized conjugation in nonsmooth nonconvex optimization, *Applied Mathematics and Optimization* (2020), doi:10.1007/s00245-020-09676-1, arXiv:1901.02746.

[6] C. Clason and T. 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.

[7] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, 2020, arXiv:2001.00216. Work in progress.

[8] J. Dardé, N. Hyvönen, T. Kuutela, and T. Valkonen, Contact adapting electrode model for electrical impedance tomography, *SIAM Journal on Applied Mathematics* 82 (2022), 427–449, doi:10.1137/21m1396125, arXiv:2102.01926.

[9] L. C. Evans, *Partial Differential Equations*, Americal Mathematical Society, 1998.

[10] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 1996.

[11] A. Griewank, Projected Hessians for Preconditioning in One-Step One-Shot Design Optimization, *Large-Scale Nonlinear Optimization* (2006), 151––171, doi:10.1007/0-387-30065-1_10.

[12] S. Günther, N. R. Gauger, and Q. Wang, Simultaneous single-step one-shot optimization with unsteady PDEs, *Journal of Computational and Applied Mathematics* 294 (2016), 12––22, doi:10.1016/j.cam.2015.07.033.

[13] A. Hamdi and A. Griewank, Reduced quasi-Newton method for simultaneous design and optimization, *Computational Optimization and Applications* 49 (2009), 521––548, doi:10.1007/s10589-009-9306-x.

[14] A. Hamdi and A. Griewank, Properties of an augmented Lagrangian for design optimization, *Optimization Methods and Software* 25 (2010), 645––664, doi:10.1080/10556780903270910.

[15] S. B. Hazra and V. Schulz, Simultaneous Pseudo-Timestepping for PDE-Model Based Optimization Problems, *BIT Numerical Mathematics* 44 (2004), 457––472, doi:10.1023/b:bitn.0000046815.96929.b8.

[16] B. He and X. Yuan, Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective, *SIAM Journal on Imaging Sciences* 5 (2012), 119–149, doi:10.1137/100814494.

[17] M. Hintermüller, K. Ito, and K. Kunisch, The primal-dual active set strategy as a semismooth Newton method, *SIAM Journal on Optimization* 13 (2002), 865–888 (2003), doi:10.1137/s1052623401383558.

[18] M. Hintermüller and G. Stadler, An Infeasible Primal-Dual Algorithm for Total Bounded Variation–Based Inf-Convolution-Type Image Restoration, *SIAM Journal on Scientific Computation* 28 (2006), 1–23.

[19] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of Advances in Design and Control, SIAM, 2008, doi:10.1137/1.9780898718614.

[20] J. Jauhiainen, P. Kuusela, A. Seppänen, and T. Valkonen, Relaxed Gauss–Newton methods with applications to electrical impedance tomography, *SIAM Journal on Imaging Sciences* 13 (2020), 1415–1445, doi:10.1137/20m1321711, arXiv:2002.08044.

[21] L. Kaland, J. C. De Los Reyes, and N. R. Gauger, One-shot methods in function space for PDE-constrained optimal control problems, *Optimization Methods and Software* 29 (2013), 376–405, doi:10.1080/10556788.2013.774397.

[22] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley Classics Library, Wiley, 1991.

[23] R. LeVeque, J., *Finite Difference Methods for Ordinary and Partial Differential Equations*, SIAM, 2007, doi:10.1137/1.9780898717839.

[24] S. Mazurenko, J. Jauhiainen, and T. Valkonen, Primal-dual block-proximal splitting for a class of non-convex problems, *Electronic Transactions on Numerical Analysis* 52 (2020), 509–552, doi:10.1553/etna_vol52s509, arXiv:1911.06284.

[25] R. Mifflin, Semismooth and semiconvex functions in constrained optimization, *SIAM Journal on Control And Optimization* 15 (1977), 959–972, doi:10.1137/0315061.

[26] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of Grundlehren der mathematischen Wissenschaften, Springer, 2006, doi:10.1007/3-540-31247-1.

[27] L. Qi and J. Sun, A nonsmooth version of Newton's method, *Mathematical Programming* 58 (1993), 353–367, doi:10.1007/bf01581275.

[28] J. Sirignano and K. Spiliopoulos, Online Adjoint Methods for Optimization of PDEs, *Applied Mathematics and Optimization* 85 (2022), doi:10.1007/s00245-022-09852-5.

[29] E. Suonperä and T. Valkonen, Linearly convergent bilevel optimization with single-step inner methods, *Computational Optimization and Applications* (2023), arXiv:2205.04862. accepted.

[30] S. Ta’asan, One Shot Methods for Optimal Control of Distributed Parameter Systems I: Finite Dimensional Control, Technical Report 91-2, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, 1991.

[31] M. Ulbrich, Semismooth Newton methods for operator equations in function spaces, *SIAM Journal on Optimization* 13 (2002), 805–842 (2003), doi:10.1137/s1052623400371569.

[32] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of MOS-SIAM Series on Optimization, SIAM, 2011, doi:10.1137/1.9781611970692.

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

[34] T. Valkonen, Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics and Optimization* 82 (2020), doi:10.1007/s00245-018-9541-6, arXiv:1703.05705.

[35] T. Vilhunen, J. P. Kaipio, P. J. Vauhkonen, T. Savolainen, and M. Vauhkonen, Simultaneous reconstruction of electrode contact impedances and internal electrical properties: I. Theory, *Meas. Sci. Technol.* 13 (2002), 1848–1854.