SINGLE-LOOP METHODS FOR BILEVEL PARAMETER LEARNING IN INVERSE IMAGING

Ensio Suonperä*   Tuomo Valkonen†

Abstract   Bilevel optimisation is used in inverse imaging problems for hyperparameter learning/identification and experimental design, for instance, to find optimal regularisation parameters and forward operators. However, computationally, the process is costly. To reduce this cost, recently so-called single-loop approaches have been introduced. On each step of an outer optimisation method, they take just a single gradient step towards the solution of the inner problem. In this paper, we flexibilise the inner algorithm to include standard methods in inverse imaging. Moreover, as we have recently shown, significant performance improvements can be obtained in PDE-constrained optimisation by interweaving the steps of conventional iterative linear system solvers with the optimisation method. We now demonstrate how the adjoint equation in bilevel problems can also benefit from such interweaving. We evaluate the performance of our approach on identifying the deconvolution kernel for image deblurring, and the subsampling operator for magnetic resonance imaging (MRI).

1 INTRODUCTION

Bilevel optimisation has recently found significant interest in inverse imaging problems, to learn or identify regularisation parameters and kernels [11, 31, 12, 3, 13, 23, 24, 25, 15, 7], as well as for experimental design, for example, to design optimal sampling patterns for magnetic resonance imaging (MRI) [38]. For our purposes, the bilevel problem involves solving

$$(1.1) \qquad \min_{\alpha \in \mathscr{A}} J(S_u(\alpha)) + R(\alpha) \quad \text{with} \quad S_u(\alpha) \in \arg\min_{u \in U} F(u; \alpha)$$

in Hilbert spaces $\mathscr{A}$ and U. Here F would be an optimisation formulation of the inverse problem of interest, and α the aforementioned parameters that we want to choose optimally. For Tikhonov-type formulations

$$(1.2) \qquad F(u; \alpha) = f(u; \alpha) + g(Ku; \alpha),$$

where f is the data fidelity term, and $g \circ K$ forms, for example, a total variation regularisation term. For instance, to identify the parameterisation of a forward operator A_α, including the kernel of a convolution operator, or the subsampling weights of a subsampled Fourier or Radon transform as in [38], we could solve the problem

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{m} \frac{1}{2} \|u_i - z_i\|^2 + \beta \|\alpha\|_1 \quad \text{where each} \quad u_i \in \arg\min_u \frac{1}{2} \|A_\alpha u - b_i\|^2 + \lambda \mathrm{TV}(u),$$

*Department of Mathematics and Statistics, University of Helsinki, Finland. ensio.suonpera@helsinki.fi, ORCID: 0000-0002-2111-0239

†MODEMAT Research Center in Mathematical Modeling and Optimization, Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland. tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

where α is the parameterisation to be identified; $u_i = [S_u(\alpha)]_i$ are the reconstructed images corresponding to the corrupted data b_i; and z_i are the ground-truth images. The outer regularisation parameter $\beta > 0$ controls the sparsity of the parameters-to-be-identified α. This regularisation could be combined with a positivity constraint. In this example, we also fix the inner regularisation parameter $\lambda > 0$, although it could also be identified.

However, bilevel optimisation can be very expensive. At the outset, it requires solving the inner problem $\min F_u$—in our domain of interest, a possibly expensive inverse problem itself—several times to find the desired optimal parameters. Therefore, in recent years, there has been a growing interest—also from machine learning, where bilevel optimisation has relevance to adversary learning [35]—in developing "single-loop" methods, that on each step of an algorithm to solve the outer problem (1.1), only take a *single step* of a conventional optimisation method towards a minimiser of the inner objective (1.2) [40, 6, 26, 33, 46, 10]. There are also similar methods that take more than one but only a fixed (small) number of inner iterations for each outer iteration [30, 29, 18, 32], while [37, 16] take an adaptive number of inner iterations.

In [28], we introduced the single-loop approach to nonsmooth PDE-constrained optimisation, such as total-variation regularised electrical impedance tomography (EIT) reconstruction, avoiding solving the PDE on each step of the optimisation method, and instead taking a single step of a conventional linear system solver (Jacobi, Gauss–Seidel, conjugate gradients) on each step of the optimisation method. This achieved significant speedups.

We now want to apply such single-loop approaches to the adjoint equations that also surface in bilevel optimisation. Moreover, compared to the aforementioned works on single-loop bilevel optimisation methods, that generally use gradient descent for the inner problem, we want to use methods that are applicable to inverse problems with nonsmooth total variation regularisation. Gradient descent only applies to smooth problems. The next step from gradient descent is forward-backward splitting, which is applicable to simple total variation regularised denoising through a dual formulation. However, when both the data term and the regularisation term involve a difficult operator, neither forward-backward is applicable [9]. A popular choice then is the primal-dual proximal splitting method (PDPS) of Chambolle and Pock [5]. We want to use (single steps of) this method for the inner problem, yet remain flexible towards other alternatives as well.

That said, due to inherent difficulties in bilevel optimisation, in this work we cannot yet allow for a nonsmooth inner problem. Indeed, due to the use of the adjoint equation, we require the inner problem to be twice differentiable. Some recent works [32, 34] avoid second-order derivatives in the algorithm through the value function reformulation, but, nevertheless, require significant second-order and other differentiability assumptions for the convergence theory.[1] This may impose a performance penalty without a corresponding benefit in applications where the (nevertheless assumed) second-order information can easily be computed, and—this is one of our principal contributions—efficiently exploited. These and, indeed, most algorithms in the literature, also do not allow a nonsmooth R in the outer problem, as our approach does. The convergence results are, moreover, for the gradient, while we are interested in iterate convergence.

In Section 2 we, therefore, introduce and prove the convergence of an abstract tracking approach to the solution of bilevel optimisation problems, based on arbitrary solvers for the inner problem and the adjoint equation that are subject to simple tracking estimates. This abstract analysis significantly simplifies upon our earlier more specific analysis in [40]. Parts of the analysis we relegate to the appendix: integral to our convergence proofs, we recall and adapt to our needs a three-point monotonicity estimate from [42] in Appendix A. Not integral to the theory, but helpful to verify the its conditions, we also discuss the regularity of the solution mapping of the inner problem in Appendix B.

---

[1]The Polyak–Łojasiewicz inequality is also, essentially, a second-order growth assumption [36].

We treat tracking estimates for the inner problem in Section 3, proving that they hold for both forward-backward splitting and the PDPS. Correspondingly, in Section 4, we then prove that standard operator splitting schemes, such as Jacobi and Gauss–Seidel splitting, satisfy the adjoint tracking property. We finish in Section 5 with numerical experiments: identifying a convolution kernel for image deconvolution, and an optimal subsampling operator for MRI. We demonstrate, in particular, significant performance improvements from a block-Gauss–Seidel scheme for the adjoint equation. We delay the derivation of details of the numerical realisation to Appendices C and D.

NOTATION AND BASIC CONCEPTS

We write $\mathbb{L}(X; Y)$ for the space of bounded linear operators between the normed spaces X and Y, and Id for the identity operator. Generally X will be Hilbert, so we can identify it with the dual X^*. We use the notation $A > B$ (resp. $A \geq B$) to indicate that $A - B$ is positive (semi-)definite We write $\langle x, y \rangle$ for an inner product, and $B(x, r)$ for a closed ball in a relevant norm $\| \cdot \|$. For self-adjoint positive semi-definite $Q \in \mathbb{L}(X; X)$ we write $\|x\|_Q := \sqrt{\langle x, x \rangle_Q} := \sqrt{\langle Qx, x \rangle}$ and $B_Q(x, r)$ for a closed ball in the norm $\| \cdot \|_Q$. For operators $p \in \mathbb{L}(X; \mathscr{A})$, we set $\|p\|_Q := \|pQ^{1/2}\|_{\mathbb{L}(X;\mathscr{A})}$. Pythagoras' *three-point identity* in Hilbert spaces then states

$$(1.3) \qquad \langle x - y, x - z \rangle_Q = \frac{1}{2}\|x - y\|_Q^2 - \frac{1}{2}\|y - z\|_Q^2 + \frac{1}{2}\|x - z\|_Q^2 \qquad \text{for all } x, y, z \in X.$$

We also extensively use Young's inequality

$$\langle x, y \rangle_Q \leq \frac{a}{2}\|x\|_Q^2 + \frac{1}{2a}\|y\|_Q^2 \qquad \text{for all } x, y \in X, \ a > 0.$$

For $G \in C^1(X)$, we write $G'(x) \in X^*$ for the Fréchet derivative at x, and $\nabla G(x) \in X$ for its Riesz presentation, i.e., the gradient. For $E \in C^1(X; Y)$, since $E'(x) \in \mathbb{L}(X; Y)$, we use the Hilbert adjoint to define $\nabla E(x) := E'(x)^* \in \mathbb{L}(Y; X)$. Then the Hessian $\nabla^2 G(x) := \nabla[\nabla G](x) \in \mathbb{L}(X; X)$. When necessary we indicate the differentiation variable with a subscript, e.g., $\nabla_u F(u, \alpha)$.

We define $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ with the usual arithmetic on $\mathbb{R}$ extended by $t + \infty = \infty$ for all $t \in \mathbb{R}$. For convex $R : X \to \overline{\mathbb{R}}$, we write dom R for the effective domain and $\partial R(x)$ for the subdifferential at x. With slight abuse of notation, we identify $\partial R(x)$ with the set of Riesz presentations of its elements. We define the *proximal map* as $\text{prox}_R(x) := \arg\min_z \frac{1}{2}\|z - x\|^2 + R(z) = (\text{Id} + \partial R)^{-1}(x)$. For $F : X \to \overline{\mathbb{R}}$, we define $F^* : X^* \to \overline{\mathbb{R}}$ for *Fenchel conjugate* (or *convex conjugate*) of F as $F^*(x^*) = \sup_{x \in X}\{\langle x^*, x \rangle - F(x)\}$.

2 AN ABSTRACT TRACKING APPROACH

In this section, we prove the convergence of our proposed abstract tracking approach to bilevel optimisation. We start in Section 2.1 by describing the problem, and then deriving in Section 2.2 the optimality conditions that we try to solve; in particular, the adjoint equation. We then introduce and discuss in Section 2.3 the abstract algorithm. We continue by stating the assumptions of the abstract method in Section 2.4, and then proving its convergence based on these assumptions 2.5. The assumptions in particular involve inner problem and adjoint tracking conditions. We verify these for several explicit algorithms in the coming Sections 3 and 4.

2.1 PROBLEM DESCRIPTION

We slightly generalise the problem formulation (1.1), and seek to solve

$$(2.1) \qquad \min_{\alpha \in \mathscr{A}} J(S_u(\alpha)) + R(\alpha) \quad \text{subject to } S_u(\alpha) \in U \text{ satisfying} \quad 0 = G(S_u(\alpha); \alpha),$$

where, on Hilbert spaces U and $\mathscr{A}$, $J : U \to \mathbb{R}$ is convex and Fréchet differentiable, $R : \mathscr{A} \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous, and $G : U \times \mathscr{A} \to U$. We call $S_u : \mathscr{A} \to U$ the solution mapping. We will generally assume it to be well-defined (and single-valued) on the effective domain of R, i.e., in $\operatorname{dom} R := \{\alpha \in \mathscr{A} \mid R(\alpha) < \infty\}$. Of all the function, we will also be making further assumptions, including differeniability assumptions, in Section 2.4

Taking $G = \nabla_u F$ for sufficiently regular F, (2.1) is a necessary condition for (1.1). If F is convex in u, this condition is also sufficient. However, we want to allow for other optimality conditions, such as the primal-dual conditions that arise from the Frenchel–Rockafellar theorem, and that our used in the context of primal-dual methods (see, e.g., [9]). We therefore allow for abstract optimality conditions for the inner problem via a general G.

2.2 OPTIMALITY CONDITIONS AND THE ADJOINT EQUATION

Assume that both G and S_u are Fréchet differentiable. Suppose a solution $S_u(\alpha)$ exists for all α near $\tilde{\alpha} \in \mathscr{A}$. Then $G(S_u(\alpha); \alpha) = 0$ for all such α, so by implicit differentiation

$$0 = \nabla_\alpha S_u(\alpha) \nabla_u G(S_u(\alpha), \alpha) + \nabla_\alpha G(S_u(\alpha), \alpha).$$

That is, $p = \nabla_\alpha S_u(\alpha)$ solves for $u = S_u(\alpha)$ the *adjoint equation*

$$(2.2) \qquad\qquad\qquad 0 = p \nabla_u G(u, \alpha) + \nabla_\alpha G(u, \alpha).$$

We introduce the corresponding solution mapping for the *adjoint variable p*,

$$(2.3) \qquad\qquad\qquad S_p(u, \alpha) := -\nabla_\alpha G(u; \alpha) \left(\nabla_u G(u; \alpha)\right)^{-1}.$$

We will later make assumptions that ensure that S_p is well-defined. Then $\nabla S_u(\alpha) = S_p(S_u(\alpha), \alpha)$.

Since $S_u : \mathscr{A} \to U$, the Fréchet derivative $S_u'(\alpha) \in \mathbb{L}(\mathscr{A}; U)$ and the Hilbert adjoint $\nabla_\alpha S_u(\alpha) \in \mathbb{L}(U; \mathscr{A})$ for all α. Consequently $p \in \mathbb{L}(U; \mathscr{A})$.

By the sum rule for Clarke subdifferentials (denoted ∂_C) and their compatibility with convex subdifferentials and Fréchet differentiable functions [8], we obtain

$$\partial_C(J \circ S_u + R)(\widehat{\alpha}) = \nabla_\alpha(J \circ S_u)(\widehat{\alpha}) + \partial R(\widehat{\alpha}) = \nabla_\alpha S_u(\widehat{\alpha}) \nabla_u J(S_u(\widehat{\alpha})) + \partial R(\widehat{\alpha}).$$

The Fermat principle for Clarke or Mordukhovich subdifferentials then furnishes the necessary optimality condition

$$(2.4) \qquad\qquad 0 \in \nabla_\alpha(J \circ S_u)(\widehat{\alpha}) + \partial R(\widehat{\alpha}) = \nabla_\alpha S_u(\widehat{\alpha}) \nabla_u J(S_u(\widehat{\alpha})) + \partial R(\widehat{\alpha}).$$

We combine the inner optimality condition $G(S_u(\alpha); \alpha) = 0$, the adjoint equation (2.2), and the outer optimality condition (2.4) as the inclusion

$$(2.5a) \qquad\qquad\qquad 0 \in H(\widehat{u}, \widehat{p}, \widehat{\alpha})$$

with

$$(2.5b) \qquad H(u, p, \alpha) := \begin{pmatrix} G(u; \alpha) \\ p\nabla_u G(u; \alpha) + \nabla_\alpha G(u; \alpha) \\ p\nabla_u J(u) + \partial R(\alpha) \end{pmatrix} \quad \text{for all} \quad u \in U, p \in \mathbb{L}(U; \mathscr{A}), \alpha \in \mathscr{A}.$$

This is the optimality condition that our proposed methods attempt to satisfy.

Algorithm 2.1 Bilevel abstract tracking approach (BATA)

Require: Functions $J : U \to \mathbb{R}$, and $R : \mathscr{A} \to \overline{\mathbb{R}}$, with J Fréchet differentiable and R convex, on Hilbert spaces U and $\mathscr{A}$. Outer step length $\sigma > 0$. Norm-defining $Q \in \mathbb{L}(U; U)$ and tracking parameters $\kappa_u, \kappa_p > 1$; $\pi_u, \pi_p > 0$; as well as $C_S > 0$, satisfying the further bounds in Assumption 2.2.

1: Pick an initial iterate $(u^0, p^0, \alpha^0) \in U \times \mathbb{L}(U; \mathscr{A}) \times \mathscr{A}$. Set $\alpha^{-1} := \alpha^0$.
2: **for** $k \in \mathbb{N}$ **do**
3:     Find (with an algorithm from Section 3) a $u^{k+1} \in U$ satisfying the inner tracking property

$$\kappa_u \|u^{k+1} - S_u(\alpha^k)\|_Q \le \|u^k - S_u(\alpha^{k-1})\|_Q + \pi_u \|\alpha^k - \alpha^{k-1}\|.$$

4:     Find (with an algorithm from Section 4) a $p^{k+1} \in \mathbb{L}(U; \mathscr{A})$ satisfying the adjoint tracking property

$$\kappa_p \|p^{k+1} - \nabla_\alpha S_u(\alpha^k)\|_{Q^{-1}} \le \|p^k - \nabla_\alpha S_u(\alpha^k)\|_{Q^{-1}} + C_S \|u^{k+1} - S_u(\alpha^k)\|_Q + \pi_p \|\alpha^k - \alpha^{k-1}\|.$$

5:     Update $\alpha^{k+1} := \mathrm{prox}_{\sigma R} \left(\alpha^k - \sigma p^{k+1} \nabla_u J(u^{k+1}) \right).$
6: **end for**

2.3 ALGORITHM

As already discussed, we generally assume that there exist solution mappings S_u and S_p of the inner problem and the adjoint equation, i.e., the first two lines of the inclusion (2.5). However, our algorithms do not attempt to solve these computationally expensive equations exactly on each step. Instead, in the abstract Algorithm 2.1, we assume to be given an abstract inner algorithm and an abstract adjoint algorithm, that satisfy a corresponding *inner tracking* and *adjoint tracking* property (Lines 3 and 4 of the algorithm). The idea is to implement those steps by taking a single step of a conventional optimisation method (Section 3), or a linear system solver (Section 4). The corresponding algorithm defines the tracking parameters, that are never to be explicitly used; they are merely needed for the convergence theory.

2.4 ASSUMPTIONS

We next state essential structural, initialisation, and step length assumptions. We start with a contractivity condition needed for the proximal step with respect to R.

Assumption 2.1. Let $R : \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. We say that R is *locally prox-σ-contractive at $\widehat{\alpha} \in \mathscr{A}$ for $q \in \mathscr{A}$* (within $A \subset \mathrm{dom}\, R$) if there exist $\sigma, C_R > 0$ and a neighbourhood $A \subset \mathrm{dom}\, R$ of $\widehat{\alpha}$ such that, for all $\alpha \in A$,

$$\|D_{\sigma R}(\alpha) - D_{\sigma R}(\widehat{\alpha})\| \le \sigma C_R \|\alpha - \widehat{\alpha}\| \quad \text{for} \quad D_{\sigma R}(\alpha) := \mathrm{prox}_{\sigma R}(\alpha - \sigma q) - \alpha.$$

If $\sigma > 0$ can be arbitrary with the same factor C_R, we drop the word "locally".

The assumption holds for smooth functions [40, Theorem A.4]. It also holds for indicator functions of convex sets if $q = 0$ [40, Theorem A.2], and $R = \beta\| \cdot \|_1 + \delta_{[0,\infty)^n}$ if $-q = (\beta, \ldots, \beta) \in \partial R(\widehat{\alpha})$ [40, Theorem A.1]. When applying the assumption to $\widehat{\alpha}$ satisfying (2.5), we will take $-q = -\widehat{p} \nabla_u J(\widehat{u}) \in \partial R(\widehat{\alpha})$. The restriction on q in the two nonsmooth examples thus serves to forbid strict complementarity: optimal solutions cannot involve interior subdifferentials. Intuitively, this restriction serves to forbid the *finite identification* property [20] of proximal-type methods, as $\{\alpha^n\}$ cannot converge too fast in our current proof techniques for the stability of the inner problem and adjoint with respect to perturbations of α.

We now come to our main assumption for the abstract algorithm.

Assumption 2.2. Let U and $\mathscr{A}$ be a Hilbert spaces. Let $R : \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, and $J : U \to \mathbb{R}$ be convex and Fréchet differentiable. Also let $G : U \times \mathscr{A} \to U$. Pick $(\widehat{u}, \widehat{p}, \widehat{\alpha}) \in H^{-1}(0)$ and let $\{(u^n, p^n, \alpha^n)\}_{n \in \mathbb{N}}$ be generated by Algorithm 2.1 for a given initial iterate $(u^0, p^0, \alpha^0) \in U \times \mathbb{L}(U; \mathscr{A}) \times \operatorname{dom} R$. For some positive $r_\alpha, r_u > 0$, we then suppose that

(i) There exists an inner problem solution mapping $S_u \in C^1(\mathscr{A}_{2r}; U)$, satisfying $G(S_u(\alpha); \alpha) = 0$ for all $\alpha \in \mathscr{A}_{2r} := B(\widehat{\alpha}, 2r_\alpha) \cap \operatorname{dom} R$. For a positive definite self-adjoint operator $Q \in \mathbb{L}(U; U)$, this mapping is also L_{S_u}-Lipschitz w.r.t. the Q-norm, i.e. $\|S_u(\alpha_1) - S_u(\alpha_2)\|_Q \le L_{S_u} \|\alpha_1 - \alpha_2\|$ for $\alpha_1, \alpha_2 \in \mathscr{A}_{2r}$.

(ii) We are given an *inner algorithm* A_u that satisfies the inner tracking property

$$\kappa_u \|A_u(u, \alpha_2) - S_u(\alpha_2)\|_Q \le \|u - S_u(\alpha_1)\|_Q + \pi_u \|\alpha_2 - \alpha_1\|$$

for any $u \in B_Q(\widehat{u}, r_u)$ and $\alpha_1, \alpha_2 \in \mathscr{A}_{2r}$ with $\pi_u > 0$ and $\kappa_u > 1$.

(iii) We are given an *adjoint algorithm* A_p that satisfies the adjoint tracking property

$$\kappa_p \|A_p(u, p, \alpha_2) - \nabla_\alpha S_u(\alpha_2)\|_{Q^{-1}} \le \|p - \nabla_\alpha S_u(\alpha_1)\|_{Q^{-1}} + C_S \|u - S_u(\alpha_2)\|_Q + \pi_p \|\alpha_2 - \alpha_1\|$$

for any $p \in \mathbb{L}(U; \mathscr{A}), u \in B_Q(\widehat{u}, r_u)$ and $\alpha_1, \alpha_2 \in \mathscr{A}_{2r}$ with $C_S, \pi_p > 0$ and $\kappa_p > 1$.

(iv) The outer fitness function J is Lipschitz continuously differentiable with factor $L_{\nabla J}$, and the function $J \circ S_u$ is γ_α-strongly convex and L_α-Lipschitz differentiable in $B(\widehat{\alpha}, r) \cap \operatorname{dom} R$ for some $\gamma_\alpha, L_\alpha > 0$. Moreover, R is locally prox-σ-contractive at $\widehat{\alpha}$ for $\widehat{p} \nabla_u J(\widehat{u})$ within $B(\widehat{\alpha}, r_\alpha) \cap \operatorname{dom} R$ for some $C_R \ge 0$.

(v) The relative initialization bounds

$$\|u^1 - S_u(\alpha^0)\|_Q \le C_u \|\alpha^0 - \widehat{\alpha}\| \quad \text{and} \quad \|p^1 - \nabla_\alpha S_u(\alpha^0)\|_{Q^{-1}} \le C_p \|\alpha^0 - \widehat{\alpha}\|$$

hold with constants $C_u, C_p > 0$, which satisfy

$$C_S \kappa_u^{-1} C_u < (\kappa_p - 1) C_p \quad \text{and} \quad C_\alpha := L_{\nabla J} N_p C_u + N_{\nabla J} C_p < \gamma_\alpha$$

for $N_p := N_{\nabla S_u} + C_p r_\alpha$,

$$N_{\nabla J} := \max_{\alpha \in B(\widehat{\alpha}, r_\alpha) \cap \operatorname{dom} R} \|\nabla_u J(S_u(\alpha))\|_Q, \quad \text{and} \quad N_{\nabla S_u} := \max_{\alpha \in B(\widehat{\alpha}, r_\alpha) \cap \operatorname{dom} R} \|\nabla_\alpha S_u(\alpha)\|_{Q^{-1}}.$$

(vi) The initial outer iterate $\alpha^0 \in B(\widehat{\alpha}, r)$ for $r = \min\{r_\alpha, r_u/(C_u + L_{S_u})\}$.

(vii) The outer step length $\sigma > 0$ satisfies

$$\sigma \le \frac{1}{C_\alpha + L_\alpha + C_R} \min \left\{ \frac{(\kappa_u - 1) C_u}{\pi_u + \kappa_u C_u}, \frac{(\kappa_p - 1) C_p - C_S \kappa_u^{-1} C_u}{\pi_p + \kappa_p C_p + C_S \kappa_u^{-1} \pi_u} \right\}$$

and

$$\sigma < (\gamma_\alpha - C_\alpha) \frac{2}{L_\alpha^2}.$$

Remark 2.3 (Interpretation). The condition (i) of Assumption 2.2 ensures that the inner problem solutions do not vary uncontrollably as the parameter α changes. This is necessary to ensure that small changes of one variable result in small changes in the other variables as well. We discuss the condition more in the next Remark 2.4.

The conditions (ii) and (iii) likewise ensure that the specific algorithms that are used to generate inner and adjoint iterates, produce small steps in response to small changes in α, and, with no change, are contractive. We verify these conditions for exemplary optimisation algorithms and linear system splitting schemes in Sections 3 and 4.

The first part of (iv) is a second order growth and boundedness condition, standard in smooth optimisation. Together with (i), the condition ensures that $\alpha \mapsto \nabla_\alpha(J \circ S_u)(\alpha)$ is Lipschitz in $B(\widehat{\alpha}, r)$. The local strong convexity of $J \circ S_u$ in the smooth case amounts to $\nabla^2_\alpha(J \circ S_u)(\alpha) > \gamma_\alpha \operatorname{Id}$ holding locally. We refer to [40, Remark 2.3] on how to reduce the latter to properties of $\nabla^2_u J(S(\alpha))$ and of S. The second part of (iv) essentially prevents $\operatorname{prox}_{\sigma R}$ from having a finite identification property, as discussed above.

The condition (v) ensures that the initial inner problem and adjoint iterates are good *relative* to the outer problem iterate. If u^1 solves the inner problem for α^0, (v) holds for any $C_u > 0$. Therefore, (v) can always be satisfied by solving the inner problem for α^0 to high accuracy. This condition *does not* require α^0 to be close to a solution $\widehat{\alpha}$ of the entire problem. We stress that *the inequality on constants in* (v) *can always be satisfied by good relative initialisation (small $C_u, C_p > 0$)*. The constants $N_{\nabla J}$ and $N_{\nabla S_u}$ require analysing the specific inner problem: how large can the solutions $u = S_u(\alpha)$ for α in $B(\hat{\alpha}, r_\alpha)$? How fast do the vary? Practically, we will not know r_α, so want to do this analysis for a large value.

The semifinal (vi) is a standard initialisation condition for local convergence. If $r > 0$ can be arbitrarily large, we obtain *global convergence*. The final (vii) is a step length restriction on the outer problem. Practically, it requires $\sigma > 0$ to be sufficiently small.

Remark 2.4 (Existence and differentiability of the solution map). The existence and differentiability of a $S_u(\alpha) = u$ such that $G(u, \alpha) = 0$ needs to be explicitly verified. When $G = \nabla_u F$, Fermat's principle proves that $\min_u F(u, \alpha)$ has a solution. That in turn, follows from lower semicontinuity and coercivity.

Suppose then that G is continuously differentiable in both variables, and that for some $u \in B(\widehat{u}, r_u)$ and $\alpha \in B(\widehat{\alpha}, 2r_\alpha) \cap \operatorname{dom} R$ we have $G(u; \alpha) = 0$ with $\nabla_u G(u; \alpha)$ invertible. Then the implicit function theorem shows the existence of a unique continuously differentiable S_u in a neighborhood of α. Such an S_u is also Lipschitz in a neighborhood of α; see, e.g., [9, Lemma 2.11]. If $\mathscr{A}$ is finite-dimensional, a compactness argument gluing together the neighborhoods then proves the continuity and Lipschitz properties of Assumption 2.2 (i). In Appendix B we discuss relaxations of these conditions to mere right-invertibility of $G_u(u; \alpha)$ (left-invertibility of $\nabla_u G(u; \alpha)$).

Remark 2.5 (Inner and adjoint algorithms). Although for visual reasons Assumption 2.2 parametrises the inner and adjoint algorithms A_u and A_p by just u, p, and α, which will in application be previous iterates, our proofs can easily handle iteration-dependent A_u and A_p, and therefore dependence on the entire history of iterates.

2.5 CONVERGENCE ANALYSIS

We now prove the convergence of Algorithm 2.1 subject to Assumption 2.2. Throughout, we take the assumption holding as a given, and *use the constants from it*. We also tacitly take it that $\alpha^n \in \operatorname{dom} R$ for all $n \in \mathbb{N}$, as this is guaranteed by the assumptions for $n = 0$, and by the proximal step in the algorithm for $n \geq 1$.

An essential goal here is to bound the error in the inner and adjoint iterates u^{n+1} and p^{n+1} in terms of the outer iterates α^n with

$$(2.6a) \qquad \|u^{n+1} - S_u(\alpha^n)\|_Q \leq C_u \|\alpha^n - \widehat{\alpha}\|, \quad \text{and}$$

$$(2.6b) \qquad \|p^{n+1} - \nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}} \leq C_p \|\alpha^n - \widehat{\alpha}\|.$$

We prove by induction that Assumption 2.2 implies these bounds for each $n \in \mathbb{N}$. We also derive bounds on the outer steps, and local monotonicity estimates.

Lemma 2.6. *Suppose Assumption 2.2 holds. Let $n \in \mathbb{N}$ and suppose that (2.6b) holds with $\alpha^n \in B(\widehat{\alpha}, r)$. Then $\|p^{n+1}\|_{Q^{-1}} \le N_p$.*

Proof. We estimate using (2.6b) and the definitions of the relevant constants in Assumption 2.2 that

$$\|p^{n+1}\|_{Q^{-1}} \le \|\nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}} + \|p^{n+1} - \nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}}$$
$$\le N_{\nabla S_u} + C_p \|\alpha^n - \widehat{\alpha}\| \le N_{\nabla S_u} + C_p r_\alpha = N_p. \qquad \square$$

The next lemma bounds the steps taken for the outer problem variable.

Lemma 2.7. *Let $n \in \mathbb{N}$. Suppose Assumption 2.2 and (2.6) hold, and $\alpha^n \in B(\widehat{\alpha}, r)$. Then*

$$(2.7) \qquad \|\alpha^{n+1} - \alpha^n\| \le \sigma[(C_\alpha + L_\alpha) + C_R]\|\alpha^n - \widehat{\alpha}\|,$$

$$(2.8) \qquad C_u \|\alpha^n - \widehat{\alpha}\| + \pi_u \|\alpha^{n+1} - \alpha^n\| \le \kappa_u C_u \big(\|\alpha^n - \widehat{\alpha}\| - \|\alpha^{n+1} - \alpha^n\|\big) \le \kappa_u C_u \|\alpha^{n+1} - \widehat{\alpha}\|$$

and

$$(2.9) \qquad (C_p + C_S \kappa_u^{-1} C_u)\|\alpha^n - \widehat{\alpha}\| + (\pi_p + C_S \kappa_u^{-1} \pi_u)\|\alpha^{n+1} - \alpha^n\| \le \kappa_p C_p \|\alpha^{n+1} - \widehat{\alpha}\|.$$

Proof. Using the α-update of Algorithm 2.1, we estimate

$$\|\alpha^{n+1} - \alpha^n\| = \|[\text{prox}_{\sigma R}(\alpha^n - \sigma p^{n+1} \nabla_u J(u^{n+1})) - \alpha^n] - [\text{prox}_{\sigma R}(\widehat{\alpha} - \sigma \widehat{p} \nabla_u J(\widehat{u})) - \widehat{\alpha}]\|$$
$$\le \|[\text{prox}_{\sigma R}(\alpha^n - \sigma p^{n+1} \nabla_u J(u^{n+1})) - \alpha^n] - [\text{prox}_{\sigma R}(\alpha^n - \sigma \widehat{p} \nabla_u J(\widehat{u})) - \alpha^n]\|$$
$$+ \|[\text{prox}_{\sigma R}(\alpha^n - \sigma \widehat{p} \nabla_u J(\widehat{u})) - \alpha^n] - [\text{prox}_{\sigma R}(\widehat{\alpha} - \sigma \widehat{p} \nabla_u J(\widehat{u})) - \widehat{\alpha}]\|.$$

Since proximal maps are 1-Lipschitz, and R is by Assumption 2.2 (iv) locally prox-σ-contractive at $\widehat{\alpha}$ for $\widehat{p} \nabla_u J(\widehat{u})$ within $B(\widehat{\alpha}, r) \cap \text{dom}\, R$ with factor C_R, it follows

$$(2.10) \qquad \|\alpha^{n+1} - \alpha^n\| \le \sigma \|p^{n+1} \nabla_u J(u^{n+1}) - \widehat{p} \nabla_u J(\widehat{u})\| + \sigma C_R \|\alpha^n - \widehat{\alpha}\| =: \sigma A + \sigma C_R \|\alpha^n - \widehat{\alpha}\|.$$

We have $\widehat{p} \nabla_u J(\widehat{u}) = \nabla_\alpha S_u(\widehat{\alpha}) \nabla_u J(S_u(\widehat{\alpha})) = \nabla_\alpha (J \circ S_u)(\widehat{\alpha})$, where $\nabla_\alpha (J \circ S_u)$ is L_α-Lipschitz in $B(\widehat{\alpha}, r) \ni \alpha^n$ by Assumption 2.2 (iv). Hence

$$A \le \|p^{n+1} \nabla_u J(u^{n+1}) - \nabla_\alpha (J \circ S_u)(\alpha^n) + \nabla_\alpha (J \circ S_u)(\alpha^n) - \widehat{p} \nabla_u J(\widehat{u})\|$$
$$\le \|p^{n+1} \nabla_u J(u^{n+1}) - \nabla_\alpha S_u(\alpha^n) \nabla_u J(S_u(\alpha^n))\| + L_\alpha \|\alpha^n - \widehat{\alpha}\|.$$

Using the Lipschitz continuity of $\nabla_u J$ from Assumption 2.2 (iv), we continue

$$A \le \|p^{n+1}(\nabla_u J(u^{n+1}) - \nabla_u J(S_u(\alpha^n))) + (p^{n+1} - \nabla_\alpha S_u(\alpha^n)) \nabla_u J(S_u(\alpha^n))\| + L_\alpha \|\alpha^n - \widehat{\alpha}\|$$
$$\le \|p^{n+1}\|_{Q^{-1}} L_{\nabla J} \|u^{n+1} - S_u(\alpha^n)\|_Q + \|p^{n+1} - \nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}} \|\nabla_u J(S_u(\alpha^n))\|_Q + L_\alpha \|\alpha^n - \widehat{\alpha}\|.$$

Since $\alpha^n \in B(\widehat{\alpha}, r)$, we have $\|p^{n+1}\|_{Q^{-1}} \le N_p$ by Lemma 2.6 and $\|\nabla_u J(S_u(\alpha^n))\|_Q \le N_{\nabla J}$ by the definition in Assumption 2.2 (v). Using (2.6) therefore gives

$$A \le N_p L_{\nabla J} C_u \|\alpha^n - \widehat{\alpha}\| + N_{\nabla J} C_p \|\alpha^n - \widehat{\alpha}\| + L_\alpha \|\alpha^n - \widehat{\alpha}\| = (C_\alpha + L_\alpha)\|\alpha^n - \widehat{\alpha}\|.$$

Inserting this into (2.10), we obtain (2.7). Assumption 2.2 (vii) and (2.7) then yield

$$(\pi_u + \kappa_u C_u)\|\alpha^{n+1} - \alpha^n\| \le \sigma(\pi_u + \kappa_u C_u)[C_\alpha + L_\alpha + C_R]\|\alpha^n - \widehat{\alpha}\| \le (\kappa_u - 1)C_u \|\alpha^n - \widehat{\alpha}\|.$$

Rearranging terms and finishing with the triangle inequality we get (2.8). We obtain (2.9) analogously to (2.8). $\qquad \square$

We discussed taking gradient steps instead of proximal steps with respect to R in [40, Remark 3.10]. We next prove that if both inner problem and adjoint iterates have small error, and we take a short step in the outer problem, then also the next inner problem and adjoint iterate has small error.

Lemma 2.8. *Let $k \in \mathbb{N}$. Suppose Assumption 2.2 holds. If $\alpha^n \in B(\widehat{\alpha}, r)$ and (2.6) holds for $n = k$, then (2.6) holds for $n = k+1$ and we have $\alpha^{k+1} \in B(\widehat{\alpha}, 2r)$.*

Proof. First, we show that $\alpha^{k+1} \in B(\widehat{\alpha}, 2r)$ and $u^{k+1} \in B_Q(\widehat{u}, r_u)$. We use (2.8) of Lemma 2.7. Its first inequality readily implies either $\|\alpha^k - \widehat{\alpha}\| > \|\alpha^{k+1} - \alpha^k\|$ or $\alpha^{k+1} = \widehat{\alpha} \in B(\widehat{\alpha}, 2r)$. In the former, using $\alpha^k \in B(\widehat{\alpha}, r)$, we get

$$\|\alpha^{k+1} - \widehat{\alpha}\| \le \|\alpha^{k+1} - \alpha^k\| + \|\alpha^k - \widehat{\alpha}\| < 2\|\alpha^k - \widehat{\alpha}\| \le 2r.$$

We estimate using (2.6a) for $n = k$ and Lipschitz continuity of S_u

$$\begin{aligned}
\|u^{k+1} - \widehat{u}\|_Q &\le \|u^{k+1} - S_u(\alpha^k)\|_Q + \|S_u(\alpha^k) - S_u(\widehat{\alpha})\|_Q \\
&\le (C_u + L_{S_u})\|\alpha^k - \widehat{\alpha}\| \le (C_u + L_{S_u})r \le r_u.
\end{aligned}$$

Thus $u^{k+1} \in B_Q(\widehat{u}, r_u)$, so that we may apply Assumption 2.2 (ii) and (iii) with $u = u^{k+1}$.

The former with (2.6a) for $n = k$ then imply

$$\begin{aligned}
(2.11) \qquad \kappa_u \|u^{k+2} - S_u(\alpha^{k+1})\|_Q &\le \|u^{k+1} - S_u(\alpha^k)\|_Q + \pi_u\|\alpha^{k+1} - \alpha^k\| \\
&\le C_u\|\alpha^k - \widehat{\alpha}\| + \pi_u\|\alpha^{k+1} - \alpha^k\|.
\end{aligned}$$

Inserting (2.8) here, we establish (2.6a) for $n = k+1$.

Similarly, Assumption 2.2 (iii), (2.11) and (2.6b) for $n = k$ imply

$$\begin{aligned}
\kappa_p \|p^{k+2} - \nabla_\alpha S_u(\alpha^{k+1})\|_{Q^{-1}} &\le \|p^{k+1} - \nabla_\alpha S_u(\alpha^k)\|_{Q^{-1}} + C_S\|u^{k+2} - S_u(\alpha^{k+1})\|_Q + \pi_p\|\alpha^{k+1} - \alpha^k\| \\
&\le (C_p + C_S\kappa_u^{-1}C_u)\|\alpha^k - \widehat{\alpha}\| + (\pi_p + C_S\kappa_u^{-1}\pi_u)\|\alpha^{k+1} - \alpha^k\|.
\end{aligned}$$

Inserting (2.9) here, we establish (2.6b) for $n = k+1$. $\qquad\square$

The next lemma is a crucial monotonicity-type estimate for the outer problem. It depends on an α-relative exactness condition on the inner and adjoint variables.

Lemma 2.9. *Let $n \in \mathbb{N}$. Suppose Assumption 2.2 (iv) and (v) hold with $\alpha^n \in B(\widehat{\alpha}, r)$ and (2.6). Then, for any $t > 0$ and $d > 0$, for some $q^{n+1} \in \partial R(\alpha^{n+1})$,*

$$\begin{aligned}
(2.12) \quad \langle p^{n+1}\nabla_u J(u^{n+1}) + q^{n+1}, \alpha^{n+1} - \widehat{\alpha} \rangle \ge -\frac{L_\alpha}{4t}\|\alpha^{n+1} - \alpha^n\|^2 \\
+ \left(\frac{\gamma_\alpha - tL_\alpha}{2} - \frac{C_\alpha}{2d}\right)\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \left(\frac{\gamma_\alpha - tL_\alpha}{2} - \frac{C_\alpha d}{2}\right)\|\alpha^n - \widehat{\alpha}\|^2.
\end{aligned}$$

Proof. The α-update of Algorithm 2.1 in implicit form reads

$$(2.13) \qquad 0 = \sigma(q^{n+1} + p^{n+1}\nabla_u J(u^{n+1})) + \alpha^{n+1} - \alpha^n \quad \text{for some} \quad q^{n+1} \in \partial R(\alpha^{n+1}).$$

Similarly, $0 \in H(\widehat{u}, \widehat{p}, \widehat{\alpha})$ implies $\widehat{p}\,\nabla_u J(\widehat{u}) + \widehat{q} = 0$ for some $\widehat{q} \in \partial R(\widehat{\alpha})$. Writing E_0 for the left hand side of (2.12), these expressions and the monotonicity of ∂R yield

$$\begin{aligned}
(2.14) \qquad E_0 &= \langle p^{n+1}\nabla_u J(u^{n+1}) - \widehat{p}\,\nabla_u J(\widehat{u}) + q^{n+1} - \widehat{q}, \alpha^{n+1} - \widehat{\alpha} \rangle \\
&\ge \langle p^{n+1}\nabla_u J(u^{n+1}) - \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n)), \alpha^{n+1} - \widehat{\alpha} \rangle \\
&\quad + \langle \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n)) - \widehat{p}\,\nabla_u J(\widehat{u}), \alpha^{n+1} - \widehat{\alpha} \rangle =: E_1 + E_2.
\end{aligned}$$

We estimate E_1 and E_2 separately. Assumption 2.2 (iv) and Theorem A.1 yield

$$(2.15) \qquad E_2 \ge \frac{\gamma_\alpha - tL_\alpha}{2}(\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \|\alpha^n - \widehat{\alpha}\|^2) - \frac{L_\alpha}{4t}\|\alpha^{n+1} - \alpha^n\|^2$$

for $t > 0$. To estimate E_1 we rearrange

$$E_1 = \langle p^{n+1}\nabla_u J(u^{n+1}) - \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n)), \alpha^{n+1} - \widehat{\alpha}\rangle$$
$$= \langle p^{n+1}(\nabla_u J(u^{n+1}) - \nabla_u J(S_u(\alpha^n))) + (p^{n+1} - \nabla_\alpha S_u(\alpha^n))(\nabla_u J(S_u(\alpha^n)), \alpha^{n+1} - \widehat{\alpha}\rangle.$$

We have $\|p^{n+1}\|_{Q^{-1}} \le N_p$ by Lemma 2.6 and $\|\nabla_u J(S_u(\alpha^n))\|_Q \le N_{\nabla J}$ by the definition of the latter in Assumption 2.2 (v) with $\alpha^n \in B(\widehat{\alpha}, r)$. The same assumptions establish that $\nabla_u J$ is Lipschitz. Hence it holds

$$E_1 \ge -\|p^{n+1}\|_{Q^{-1}}\|\nabla_u J(u^{n+1}) - \nabla_u J(S_u(\alpha^n))\|_Q\|\alpha^{n+1} - \widehat{\alpha}\|$$
$$- \|p^{n+1} - \nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}}\|\nabla_u J(S_u(\alpha^n))\|_Q\|\alpha^{n+1} - \widehat{\alpha}\|$$
$$\ge -\left(L_{\nabla J}N_p\|u^{n+1} - S_u(\alpha^n)\|_Q + N_{\nabla J}\|p^{n+1} - \nabla_\alpha S_u(\alpha^n)\|_{Q^{-1}}\right)\|\alpha^{n+1} - \widehat{\alpha}\|.$$

Applying (2.6) and Young's inequality now yields for any $d > 0$ the estimate

$$(2.16) \qquad E_1 \ge -\left(L_{\nabla J}N_p C_u + C_p N_{\nabla J}\right)\|\alpha^n - \widehat{\alpha}\|\|\alpha^{n+1} - \widehat{\alpha}\|$$
$$\ge -\left(L_{\nabla J}N_p C_u + C_p N_{\nabla J}\right)\left(\frac{d}{2}\|\alpha^n - \widehat{\alpha}\|^2 + \frac{1}{2d}\|\alpha^{n+1} - \widehat{\alpha}\|^2\right).$$

By inserting (2.15) and (2.16) into (2.14) we obtain the claim (2.12). □

By combining the previous lemmas, we prove an estimate from which local convergence is immediate.

Lemma 2.10. *Suppose Assumption 2.2 and the inner and adjoint exactness estimate (2.6) hold with $\alpha^n \in B(\widehat{\alpha}, r)$ for a $n \in \mathbb{N}$. Then*

$$(2.17) \qquad (1 + \sigma\varepsilon_\sigma)\|\alpha^{n+1} - \widehat{\alpha}\|^2 \le \|\alpha^n - \widehat{\alpha}\|^2$$

for

$$\varepsilon_\sigma := \gamma_\alpha - \sigma L_\alpha^2/2 - C_\alpha^2(\gamma_\alpha - \sigma L_\alpha^2/2)^{-1} > 0.$$

Proof. We observe that $\varepsilon_\sigma > 0$ by the inequality in Assumption 2.2 (v). Since $\alpha^n \in B(\widehat{\alpha}, r)$ and (2.6) hold, Lemma 2.9 with $t = \sigma L_\alpha/2$ gives the monotonicity estimate

$$\langle \sigma\left(p^{n+1}\nabla_u J(u^{n+1}) + q^{n+1}\right), \alpha^{n+1} - \widehat{\alpha}\rangle \ge$$
$$-\frac{1}{2}\|\alpha^{n+1} - \alpha^n\|^2 + \frac{\sigma(\gamma_\alpha - \sigma L_\alpha^2/2 - C_\alpha d^{-1})}{2}\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \frac{\sigma(\gamma_\alpha - \sigma L_\alpha^2/2 - C_\alpha d)}{2}\|\alpha^n - \widehat{\alpha}\|^2$$

for some $q^{n+1} \in \partial R(\alpha^{n+1})$ and a constant $d > 0$. Choosing $d = C_\alpha^{-1}(\gamma_\alpha - \sigma L_\alpha^2/2)$, we get

$$\gamma_\alpha - \sigma L_\alpha^2/2 - C_\alpha d^{-1} = \varepsilon_\sigma \quad \text{and} \quad \gamma_\alpha - \sigma L_\alpha^2/2 - C_\alpha d = 0.$$

It follows

$$(2.18) \qquad \langle \sigma\left(p^{n+1}\nabla_u J(u^{n+1}) + q^{n+1}\right), \alpha^{n+1} - \widehat{\alpha}\rangle \ge -\frac{1}{2}\|\alpha^{n+1} - \alpha^n\|^2 + \frac{\sigma\varepsilon_\sigma}{2}\|\alpha^{n+1} - \widehat{\alpha}\|^2.$$

We now come to the fundamental argument of the testing approach of [42], combining operator-relative monotonicity estimates with the three-point identity. Indeed, (2.18) combined with the implicit algorithm

$$0 = \sigma(q^{n+1} + p^{n+1}\nabla_u J(u^{n+1})) + \alpha^{n+1} - \alpha^n \quad \text{for some} \quad q^{n+1} \in \partial R(\alpha^{n+1})$$

gives

$$\langle \alpha^{n+1} - \alpha^n, \alpha^{n+1} - \widehat{\alpha}\rangle + \frac{\sigma\varepsilon_\sigma}{2}\|\alpha^{n+1} - \widehat{\alpha}\|^2 \le \frac{1}{2}\|\alpha^{n+1} - \alpha^n\|^2.$$

Inserting the three-point identity (1.3) yields (2.17). □

We simplify the assumptions of the previous lemma to just Assumption 2.2.

Lemma 2.11. *Suppose Assumption 2.2 holds. Then* (2.17) *holds for any* $n \in \mathbb{N}$.

Proof. The claim readily follows from Lemma 2.10 if we prove by induction for all $n \in \mathbb{N}$ that

$$(2.19) \qquad\qquad \alpha^n \in B(\widehat{\alpha}, r), \ (2.6), \ \text{and} \ (2.17) \ \text{hold.}$$

We first prove (2.19) for $n = 0$. Assumption 2.2 (v) directly establishes (2.6), and Assumption 2.2 (vi) establishes $\alpha^n \in B(\widehat{\alpha}, r)$. Now Lemma 2.10 proves (2.17) for $n = 0$. This concludes the proof of induction base.

We then make the induction assumption that (2.19) holds for $n = k$ and prove it for $n = k + 1$. The induction assumption and Lemma 2.8 give (2.6) for $n = k + 1$. The inequality (2.17) for $n = k$ and $\alpha^k \in B(\widehat{\alpha}, r)$ also ensure $\|\alpha^{k+1} - \widehat{\alpha}\| \leq \|\alpha^k - \widehat{\alpha}\| \leq r$. Now Lemma 2.10 shows (2.17) and concludes the proof of (2.19) for $n = k + 1$. $\qquad\square$

We finally come to the main convergence result for the abstract Algorithm 2.1.

Theorem 2.12. *Suppose Assumption 2.2 holds. Then for $\{\alpha^n\}_{n \in \mathbb{N}}$ generated by Algorithm 2.1, we have* $\|\alpha^n - \widehat{\alpha}\|^2 \to 0$ *linearly.*

Proof. Lemma 2.11 proves (2.17) for all $n \in \mathbb{N}$. Since, by the very same lemma, $1 + \sigma\varepsilon_\sigma > 1$ therein, the claim follows. $\qquad\square$

Corollary 2.13. *Suppose Assumption 2.2 holds. Then for $\{u^n\}_{n \in \mathbb{N}}$ generated by Algorithm 2.1, we have* $\|u^n - \widehat{u}\|^2 \to 0$ *linearly.*

Proof. We apply Theorem 2.12 and then use (2.6a) and Lipschitz continuity of S_u in

$$\|u^{n+1} - \widehat{u}\|_Q \leq \|u^{n+1} - S_u(\alpha^n)\|_Q + \|S_u(\alpha^n) - S_u(\widehat{\alpha})\|_Q \leq (C_u + L_{S_u})\|\alpha^n - \widehat{\alpha}\|. \qquad\square$$

3 SPLITTING METHODS FOR THE INNER PROBLEM

In Section 3.1 we present a general framework for optimisation algorithms that can be used for solving the inner problem. We also give examples of algorithms following the framework. In Section 3.2 we first prove inner tracking estimate for such general algorithm under appropriate assumptions. Last, we show that the aforementioned example algorithms satisfy these assumptions and thus also the inner tracking property.

3.1 A GENERAL APPROACH

Let U and $\mathscr{A}$ be Hilbert spaces, and $G = T + W$ for some $T, W : U \times \mathscr{A} \to U$. Then $u = S_u(\alpha)$ solves the inner problem if

$$(3.1) \qquad\qquad 0 = T(u; \alpha) + W(u; \alpha).$$

We assume that such a solution exists for all $\alpha \in \mathscr{A}_{2r}$, see Remark 2.4.[2] We seek to move towards $S_u(\alpha)$ with preconditioned forward-backward type algorithms that, for fixed α, given the previous iterate u^k, solve the next iterate u^{k+1} from

$$(3.2) \qquad\qquad 0 = T(u^{k+1}; \alpha) + W(u^k; \alpha) + Q(u^{k+1} - u^k)$$

for some self-adjoint preconditioning operator $Q \in \mathbb{L}(U; U)$.

---

[2]We do not here at the outset assume unique solutions, although we will later, in practise, need to impose uniqueness. In principle, for Section 2, sufficient regularity of some solution map S_u and a corresponding tracking inequality is enough.

Example 3.1 (Forward-backward splitting). Let U and $\mathscr{A}$ be Hilbert spaces. Let further $f, g : U \times \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper and lower semicontinuous in their first parameter, f also differentiable with respect to its first parameter. Then (3.1) with $T = \nabla_u f$ and $W = \partial_u g$ is a necessary and sufficient optimality condition for solution of $\min_{u \in U} f(u; \alpha) + g(u; \alpha)$. Choosing $Q = \tau^{-1} \mathrm{Id}$, the implicit form (3.2) corresponds to the forward-backward splitting

$$(3.3) \qquad u^{k+1} = \mathrm{prox}_{\tau g(\cdot\,; \alpha)}(u^k - \tau \nabla f(u^k; \alpha))$$

Example 3.2 (Primal-dual proximal splitting, PDPS). Let X, Y and $\mathscr{A}$ be Hilbert spaces, $f_0, e : X \times \mathscr{A} \to \overline{\mathbb{R}}$ and $g : Y \times \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous in their first parameter, e with an L-Lipschitz gradient with respect to its first parameter. Define $f := f_0 + e$, suppose $K \in \mathbb{L}(X; Y)$, and consider the problem

$$\min_x f(x, \alpha) + g(Kx; \alpha).$$

This can be equivalently written as the saddle point problem

$$\min_x \max_y f(x; \alpha) + \langle Kx, y \rangle - g^*(y; \alpha).$$

Following [21, 9, Theorem 5.11] and writing $u = (x, y)$, this problem has primal-dual optimality conditions of the form (3.1) with

$$(3.4) \qquad T(u; \alpha) := \begin{pmatrix} \partial_x f_0(x; \alpha) + K^* y \\ \omega^{-1}[\partial_y g^*(y; \alpha) - Kx] \end{pmatrix} \quad \text{and} \quad W(u; \alpha) := \begin{pmatrix} \nabla_x e(x; \alpha) \\ 0 \end{pmatrix}$$

for any $\omega > 0$. Defining for step length parameters $\tau_x, \tau_y > 0$ the preconditioning operator

$$(3.5) \qquad Q := \begin{pmatrix} \tau_x^{-1} \mathrm{Id} & -K^* \\ -K & \omega^{-1} \tau_y^{-1} \mathrm{Id} \end{pmatrix},$$

(3.2) then becomes an implicit form of the primal-dual proximal splitting (PDPS) method of [5], here with an additional forward step with respect to e:

$$(3.6) \qquad \begin{cases} x^{k+1} &= \mathrm{prox}_{\tau_x f_0(\cdot\,; \alpha)}(x^k - \tau_x(K^* y^k + \nabla_x e(x^k; \alpha)), \\ y^{k+1} &= \mathrm{prox}_{\tau_y g^*(\cdot\,; \alpha)}(y^k + \tau_y K((1 + \omega)x^{k+1} - \omega x^k)). \end{cases}$$

The method converges (for fixed α) if $\tau_x L/2 + \tau_x \tau_y \|K\|^2 \leq 1$ and $\omega = 1$, or is chosen to appropriately reflect available strong convexity; see, e.g., [9].

Further algorithms fitting the general framework (3.2) can be found in [9, 42].

3.2 TRACKING ESTIMATES

We now prove inner problem tracking estimates for the above methods. To do so, we start with a result for abstract methods of the form (3.2). It requires introducing abstract forms of monotonicity. Let X be a Hilbert space and $\tilde{X} \subset X$. We say that $T : X \to X$ is (Γ_T-strongly) monotone for some self-adjoint and positive semi-definite $\Gamma_T \in \mathbb{L}(X; X)$ at $\widehat{x} \in \tilde{X}$ if

$$(3.7) \qquad \langle T(x) - T(\widehat{x}), x - \widehat{x} \rangle \geq \|x - \widehat{x}\|_{\Gamma_T}^2 \quad (x \in \tilde{X}).$$

If this holds for all $\widehat{x} \in \tilde{X}$, we say that T is (Γ_T-strongly) monotone in $\tilde{X}$. Moreover, we say that $W : X \to X$ is (Γ_W-strongly) three-point monotone at $\widehat{x} \in \tilde{X}$ with respect to a self-adjoint and positive

semi-definite $\Gamma_W, \Lambda \in \mathbb{L}(X; X)$ if

$$(3.8) \qquad \langle W(z) - W(\widehat{x}), x - \widehat{x} \rangle \geq \|x - \widehat{x}\|_{\Gamma_W}^2 - \frac{1}{4}\|x - z\|_\Lambda^2 \quad (x, z \in \tilde{X}).$$

If this holds for all $\widehat{x} \in \tilde{X}$, we say that W is (Γ_W-strongly) three-point monotone with respect to Λ in $\tilde{X}$. When $\Gamma_T = \gamma\,\mathrm{Id}$, $\Gamma_P = 0$ and $\Lambda = L\,\mathrm{Id}$, these operator-relative properties reduce to standard strong and three-point monotonicity, see [9, Cororolary 7.2].

We now prove the promised tracking estimate result for general splitting methods for the inner problem. The Lipschitz properties required from S_u, were discussed in Remark 2.4.

Theorem 3.3. *Let U and $\mathscr{A}$ be Hilbert spaces, $u^k \in \tilde{U} \subset U$, $\alpha, \check{\alpha} \in \tilde{\mathscr{A}} \subset \mathscr{A}$, and operators $T, W : U \to U$. Let further $Q, \Gamma := \Gamma_T + \Gamma_W, \Lambda \in \mathbb{L}(U; U)$ be self-adjoint and positive semi-definite, and suppose that T is (Γ_T-strongly) monotone in $\tilde{U}$ and W is (Γ_W-strongly) three-point monotone with respect to Λ in $\tilde{U}$. Moreover, assume that $u = S_u(\alpha)$ that satisfies (3.1) is unique, and L_{S_u}-Lipschitz in $\tilde{\mathscr{A}}$. If*

$$(3.9a) \qquad\qquad Q + 2\Gamma \geq cQ \quad \textit{for some} \quad c > 1, \quad \textit{and}$$

$$(3.9b) \qquad\qquad Q \geq \Lambda/2,$$

then u^{k+1} generated by (3.2) satisfies Assumption 2.2 (ii) with $\kappa_u = \sqrt{c}$ and $\pi_u = L_{S_u}$, i.e.

$$\sqrt{c}\|u^{k+1} - S_u(\alpha)\|_Q \leq \|u^k - S_u(\check{\alpha})\|_Q + L_{S_u}\|\alpha - \check{\alpha}\|.$$

Proof. By the assumed monotonicity properties of T and W, we have

$$(3.10) \qquad \langle T(u^{k+1}; \alpha) + W(u^k; \alpha), u^{k+1} - S_u(\alpha) \rangle \geq \|u^{k+1} - S_u(\alpha)\|_\Gamma^2 - \frac{1}{4}\|u^{k+1} - u^k\|_\Lambda^2.$$

Inserting (3.2) into (3.10), and applying the three-point inequality (1.3) yields

$$\frac{1}{2}\|u^{k+1} - S_u(\alpha)\|_{Q+2\Gamma}^2 + \frac{1}{2}\|u^{k+1} - u^k\|_{Q-\Lambda/2}^2 \leq \frac{1}{2}\|u^k - S_u(\alpha)\|_Q^2.$$

Using (3.9) we obtain

$$\kappa_u\|u^{k+1} - S_u(\alpha)\|_Q \leq \|u^k - S_u(\alpha)\|_Q.$$

Therefore triangle inequality and Lipschitz continuity of S_u in $\tilde{\mathscr{A}}$ gives

$$\kappa_u\|u^{k+1} - S_u(\alpha)\|_Q \leq \|u^k - S_u(\check{\alpha})\|_Q + L_{S_u}\|\alpha - \check{\alpha}\|. \qquad\qquad \square$$

The next two theorems specialise Theorem 3.3 to forward-backward splitting and the PDPS.

Theorem 3.4. *Let U and $\mathscr{A}$ be Hilbert spaces with $\tilde{U} \subset U$ and $\check{\alpha} \in \tilde{\mathscr{A}} \subset \mathscr{A}$. Let further $f : U \times \mathscr{A} \to \mathbb{R}$ and $g : U \times \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous in their first argument. Suppose that f is differentiable, ∇f is L-Lipschitz continuous, and g is γ-strongly convex in $\tilde{U}$ for some $\gamma > 0$. Moreover, assume that S_u satisfying (3.1) with $T = \nabla_u f$ and $W = \partial_u g$ is single-valued and L_{S_u}-Lipschitz in $\tilde{\mathscr{A}}$. If the step length parameter $\tau > 0$ satisfies $\tau L \leq 2$, then u^{k+1} generated by forward-backward algorithm (3.3) for given $u^k \in \tilde{U}$ and $\alpha \in \tilde{\mathscr{A}}$ satisfies inner tracking property.*

Proof. This proof is based on Theorem 3.3 and to use it we need to prove (3.9). By assumption, ∂g is γ-strongly monotone, and ∇f is three-point monotone with the factor L. Thus the condition (3.9) now reads $1 + 2\gamma > 1, 1 \geq \tau L/2$ and $1 > 0$. These hold due to $\gamma > 0$ and the step length condition $\tau L \leq 2$. $\quad\square$

Remark 3.5. Similar proof works for strongly convex f and convex g. Gradient descent is a special case of this with g as the zero function.

Theorem 3.6. *Let $U = (X \times Y)$ and $\mathscr{A}$ be Hilbert spaces with $\tilde{U} = (\tilde{X} \times \tilde{Y}) \subset U$ and $\breve{\alpha} \in \tilde{\mathscr{A}} \subset \mathscr{A}$. Let further $f_0, e : X \times \mathscr{A} \to \overline{\mathbb{R}}$ and $g : Y \times \mathscr{A} \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous in their first argument, e with a L−Lipschitz gradient with respect to its first argument. Moreover, suppose f_0 is γ_f-strongly convex in $\tilde{X}$ and g^* is γ_{g^*}-strongly convex in $\tilde{Y}$ for some $\gamma_f, \gamma_{g^*} > 0$. Also let $K \in \mathbb{L}(X; Y)$, and assume that S_u satisfying (3.1) with (3.4) is single-valued and L_{S_u}-Lipschitz in $\tilde{\mathscr{A}}$. If the step length parameters $\tau_x > 0$ and $\tau_y > 0$ satisfy $\tau_x L/2 + \tau_x \tau_y \|K\|^2 \leq 1$, then $u^{k+1} = (x^{k+1}, y^{k+1})$ generated by the PDPS (3.6) for given $u^k = (x^k, y^k) \in \tilde{U}$ and $\alpha \in \tilde{\mathscr{A}}$ satisfies the inner tracking property.*

Proof. The PDPS can be written as (3.2) with operators T and W as in (3.4), and preconditioning operator Q defined in (3.5). The operator T is Γ-strongly monotone in $\tilde{U}$ and W is three-point monotone with respect to Λ in $\tilde{U}$ for

$$\Gamma := \begin{pmatrix} \gamma_f \operatorname{Id} & (1 - \omega^{-1})K^*/2 \\ (1 - \omega^{-1})K/2 & \omega^{-1}\gamma_{g^*} \operatorname{Id} \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} L \operatorname{Id} & 0 \\ 0 & 0 \end{pmatrix}.$$

The claim follows from Theorem 3.3 if we show (3.9). We first expand (3.9a) as

$$\begin{pmatrix} ((1-c)\tau_x^{-1} + 2\gamma_f) \operatorname{Id} & [(1 - \omega^{-1}) - (1 - c)]K^* \\ [(1 - \omega^{-1}) - (1 - c)]K & \omega^{-1}((1-c)\tau_y^{-1} + 2\gamma_{g^*}) \operatorname{Id} \end{pmatrix} \geq 0.$$

This readily holds by taking $\omega^{-1} = c = 1 + 2\min\{\tau_x\gamma_f, \tau_y\gamma_{g^*}\}$. To prove (3.9b), we use Young's inequality to estimate for any $\delta > 0$ that

$$2\langle x, K^*y \rangle \leq \tau_x^{-1}(1 - \delta)\|x\|^2 + \tau_x(1 - \delta)^{-1}\|K^*y\|^2 \quad \text{for all} \quad x, y.$$

Thus

$$Q = \begin{pmatrix} \tau_x^{-1} \operatorname{Id} & -K^* \\ -K & \omega^{-1}\tau_y^{-1} \operatorname{Id} \end{pmatrix} \geq \begin{pmatrix} \delta\tau_x^{-1} \operatorname{Id} & 0 \\ 0 & \omega^{-1}\tau_y^{-1} \operatorname{Id} - \tau_x(1 - \delta)^{-1}KK^* \end{pmatrix}.$$

Consequently (3.9b) holds if $\delta \geq \tau_x L/2$ and $(1 - \delta)\omega^{-1} > \tau_x\tau_y\|K\|^2$. Since $\omega^{-1} > 1$, the latter holds if $1 - \delta \geq \tau_x\tau_y\|K\|^2$. It remains to take $\delta = \tau_x L/2$, and use our assumption $\tau_x L/2 + \tau_x\tau_y\|K\|^2 \leq 1$. Note that, since $\omega^{-1} > 1$, this also proves that Q is positive definite. □

4 SPLITTING METHODS FOR THE ADJOINT

We now develop splitting methods for solving the adjoint equation based on conventional iterative splitting methods for linear systems. In Section 4.1 we present the overall approach and prove a perturbed contractivity property. We also provide example of splittings that satisfy the relevant conditions. In Section 4.2 we then prove the adjoint tracking estimate based on the perturbed contractivity property.

4.1 OPERATOR SPLITTING METHODS FOR LINEAR SYSTEMS

Let U and $\mathscr{A}$ be Hilbert spaces. We need to find $p^{k+1} \in \mathbb{L}(U; \mathscr{A})$ that approximately solves the adjoint equation

$$(4.1) \qquad\qquad p^{k+1}\nabla_u G(u^{k+1}, \alpha^k) + \nabla_\alpha G(u^{k+1}, \alpha^k) = 0$$

to such a precision that adjoint tracking property Assumption 2.2 (iii) holds. Dropping the iteration indices for brevity, this is a linear equation of the form

$$(4.2) \qquad\qquad A_v p = b_v$$

with

(4.3) $$v = (u, \alpha), \quad A_v p := p\nabla_u G(u, \alpha), \quad \text{and} \quad b_v = -\nabla_\alpha G(u, \alpha).$$

We write this section in general Hilbert spaces V and P, which in the overall setting of this paper equal $U \times \mathscr{A}$ and $\mathbb{L}(U; \mathscr{A})$. For some $v \in D \subset V$ and $A_v \in \mathbb{L}(P; P)$, let us thus be given the splitting

(4.4) $$A_v = N_v + M_v$$

with N_v invertible. Given the previous iterate $p^k \in P$, a linear system splitting method takes a step towards solution of equation (4.2) by solving p^{k+1} from

(4.5) $$N_v p^{k+1} + M_v p^k = b_v.$$

The next lemma is adapted from [28], where squared tracking estimates were needed for PDE-constrained optimisation. We need non-squared estimates. The condition (4.6) is a Lipschitz condition on A_v, which for (4.1) becomes a Lipschitz condition on $\nabla_u G$; see (4.3). The condition (4.7) is a standard contractivity condition on the splitting scheme; see [19, Theorem 10.1.1]. It obviously holds for the trivial splitting $N_v = A_v$ and $M_v = 0$ if A_v is positive definite. For (4.1) this again means that $\nabla_u G$ is positive definite. We provides examples of other splitting schemes after the lemma.

Lemma 4.1. *Let V and P be Hilbert spaces, and $A_v \in \mathbb{L}(P; P)$ for all $v \in D \subset V$. Suppose $v \mapsto b_v : V \to P$ is L_b-Lipschitz, and for some $L_A \geq 0$ that*

(4.6) $$\|A_v - A_{\tilde{v}}\| \leq L_A \|v - \tilde{v}\| \quad (v, \tilde{v} \in D).$$

Let us be given a splitting (4.4) with N_v invertible, and such that

(4.7) $$\|N_v^{-1} M_v\| \leq \zeta \quad \text{and} \quad \gamma_N \|N_v^{-1}\| \leq 1$$

for some $\zeta \in [0, 1)$ and $\gamma_N > 0$. Suppose $\check{p}$ solves $N_v \check{p} + M_v p = b_v$ for a given $p \in P$. Then for any $v, \bar{v} \in D$ and $\bar{p} \in P$ satisfying

$$A_{\bar{v}} \bar{p} = b_{\bar{v}},$$

we have

(4.8) $$\zeta \|p - \bar{p}\| \geq \|\check{p} - \bar{p}\| - (L_A \|\bar{p}\| + L_b) \gamma_N^{-1} \|v - \bar{v}\|.$$

Proof. Using (4.5) with $A_{\bar{v}} \bar{p} = b_{\bar{v}}$ and $A_v \bar{p} = N_v \bar{p} + M_v \bar{p}$, we expand

$$\begin{aligned}
\check{p} - \bar{p} &= N_v^{-1}(b_v - M_v p) - \bar{p} \\
&= N_v^{-1}[b_v - b_{\bar{v}}] + N_v^{-1}(A_{\bar{v}} - A_v)\bar{p} - N_v^{-1} M_v(p - \bar{p}).
\end{aligned}$$

Taking the norm and using the triangle inequality, the Lipschitz assumption on b, as well as (4.6), therefore

$$\|\check{p} - \bar{p}\| \leq \|N_v^{-1}\|(L_A \|\bar{p}\| + L_b)\|\bar{v} - v\| + \|N_v^{-1} M_v\|\|p - \bar{p}\|.$$

Further using (4.7), we obtain (4.8). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

If the parameters are not perturbed, we immediately get linear convergence:

Corollary 4.2. *If $v = \bar{v}$ in Lemma 4.1, then $\|\check{p} - \bar{p}\| \leq \zeta \|p - \bar{p}\|$.*

We next present example slitting methods, and conditions that guarantee (4.7).

Example 4.3 (Jacobi splitting). If N_v is the diagonal of a matrix $A_v \in \mathbb{R}^{n \times n}$, we obtain Jacobi splitting. The first part of (4.7) reduces to strict diagonal dominance, see [19, §10.1]. The second part always holds and N_v is invertible when the main diagonal of A_v has only positive entries. Then γ_N is the minimum of the diagonal values.

Example 4.4 (Gauss–Seidel splitting). If N_v is the lower triangle and diagonal of a matrix $A_v \in \mathbb{R}^{n \times n}$, we obtain Gauss–Seidel splitting. The first inequality of (4.7) holds for some $\zeta \in [0, 1)$ when A_v is symmetric and positive definite. This follows similarly to [19, proof of Theorem 10.1.2]. The second inequality holds for some γ_N when N_v is invertible, i.e., has no zeros on the main diagonal.

Example 4.5 (Identity splitting). If $\gamma \,\mathrm{Id} \leq A_v$ and $\|A_v\| \leq L$ for $\gamma, L > 0$, and $N_v = \theta^{-1} \mathrm{Id}$ for $0 < \theta < \min\{\gamma_N^{-1}, 2\gamma/L^2\}$, then (4.7) holds for $\zeta = \sqrt{1 + \theta^2 L^2 - 2\theta\gamma} \in [0, 1)$ and (any) $\gamma_N > 0$. Indeed,

$$\|N_v^{-1} M_v\|^2 = \sup_{x, \|x\| \leq 1} \langle (\mathrm{Id} - \theta A_v)x, (\mathrm{Id} - \theta A_v)x \rangle \leq 1 + \theta^2 \|A\|^2 - 2\theta \inf_{x, \|x\| \leq 1} \langle Ax, x \rangle \leq 1 + \theta^2 L^2 - 2\theta\gamma.$$

Example 4.6 (No splitting). If $N_v = A_v$, (4.7) holds with $\zeta = 0$ and γ_N the minimal eigenvalue of A_v, assumed symmetric positive definite.

Example 4.7 (Block Gauss–Seidel). Let

$$A_v = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad N_v = \begin{pmatrix} N_{11} & 0 \\ A_{21} & N_{22} \end{pmatrix}, \quad \text{and} \quad M_v = \begin{pmatrix} M_{11} & A_{12} \\ 0 & M_{22} \end{pmatrix}$$

with $A_{11} = N_{11} + M_{11}$ and $A_{22} = N_{22} + M_{22}$ for some invertible and bounded operators N_{11} and N_{22} invertible and bounded. Suppose that γ_N and $\zeta \in [0, 1)$ satisfy

$$(4.9) \qquad\qquad 0 < \gamma_N \leq \frac{\|N_{11}\| \|N_{22}\|}{2\|N_{11}\| + \|N_{22}\|(1 + \|N_{22}^{-1} A_{21}\|^2)},$$

and

$$\|N_{22}^{-1}(M_{22} - A_{21} N_{11}^{-1} A_{12})\|^2 + \|N_{22}^{-1}(M_{22} - A_{21} N_{11}^{-1} A_{12})\| \|N_{22}^{-1} A_{21}\| \|N_{11}^{-1} M_{11}\|$$
$$+ \|N_{11}^{-1} M_{11}\|^2 (1 + \|N_{11}^{-1} A_{12}\|)(1 + \|N_{22}^{-1} A_{21}\|^2) + \|N_{11}^{-1} A_{12}\|^2 + \|N_{11}^{-1} A_{12}\| \|N_{11}^{-1} M_{11}\| \leq \zeta^2.$$

Then (4.7) holds, as we prove in Appendix C. In particular, if $M_{11} = 0$ and $M_{22} = 0$, the condition on ζ reduces to the much simpler

$$(4.10) \qquad\qquad \|A_{22}^{-1} A_{21} A_{11}^{-1} A_{12}\|^2 + \|A_{11}^{-1} A_{12}\|^2 \leq \zeta^2.$$

This holds, in particular, when the diagonal blocks of A_v are invertible and large compared to the off-diagonal blocks. Moreover, there exists γ_N satisfying (4.9) when both N_{11} and N_{22} are non-zero.

Example 4.8. Let $G = T + W$ as in Example 3.2 on the PDPS for the inner problem. Assume that the functions f and g^* therein are twice continuously differentiable in their first parameter and $\nabla_x f$ and $\nabla_y g^*$ are also Lipschitz differentiable in their second parameter. We may then apply the block Gauss–Seidel splitting of Example 4.7 to the corresponding adjoint equation, which now reads

$$\begin{pmatrix} \nabla_x^2 f(x; \alpha) & K^* \\ -K & \nabla_y^2 g^*(y; \alpha) \end{pmatrix} \begin{pmatrix} p_x \\ p_y \end{pmatrix} + \begin{pmatrix} \nabla_\alpha \nabla_x f(x; \alpha) \\ \nabla_\alpha \nabla_y g^*(y; \alpha) \end{pmatrix} = 0$$

for the unknown $p = (p_x, p_y) \in P$. The simplified condition (4.10) holds if $\nabla_x^2 f(x; \alpha)$ and $\nabla_y^2 g^*(y; \alpha)$ are non-singular, and K is small compared to them. In practise, to ensure invertibility and the stability of inversion we can take N_{11} as perturbed version of $\nabla_x^2 f(x; \alpha)$. We also take $N_{22} = \nabla_y^2 g^*(y; \alpha) + \theta_y^{-1} \operatorname{Id}$ for $\theta_y > 0$, which imply the choice of M_v via $M_v = A_v - N_v$.

4.2 TRACKING ESTIMATES

We next show adjoint equation tracking estimates for the above operator splitting methods for linear systems as α varies. Following result relies heavily on Lemma 4.1.

Theorem 4.9. Let U and $\mathscr{A}$ be Hilbert spaces. Let further $G : U \times \mathscr{A} \to U$ and $S_u : \mathscr{A} \to U$ be Lipschitz continuously differentiable, and satisfy

$$\nabla_\alpha S_u(\alpha) \nabla_u G(S_u(\alpha); \alpha) + \nabla_\alpha G(S_u(\alpha); \alpha) = 0.$$

Also let $Q \in \mathbb{L}(U; U)$ be positive definite and self-adjoint. For $v = (u, \alpha) \in U \times \mathscr{A}$, define $A_v \in \mathbb{L}(P; P)$ for $P = \mathbb{L}(U; \mathscr{A})$ and b_v via

$$A_v p := p \nabla_u G(u; \alpha) \quad and \quad b_v = -\nabla_\alpha G(u; \alpha).$$

Let the splitting $N_v + M_v = A_v$ satisfy the conditions of Lemma 4.1 for some $\zeta \in [0, 1)$ and $\gamma_N > 0$. Then p^{k+1} solving (4.5) for given $p^k \in P$ satisfies the adjoint tracking property, i.e.

$$\kappa_p \| p^{k+1} - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} \leq \| p^k - \nabla_\alpha S_u(\check{\alpha}) \|_{Q^{-1}} + C_S \| u - S_u(\alpha) \|_Q + \pi_p \| \alpha - \check{\alpha} \|$$

for any $\check{\alpha} \in V_\alpha \subset \mathscr{A}$ as well as the constants

$$\kappa_p = \zeta^{-1}, \quad C_S = \sup_{\alpha \in V_\alpha} \zeta^{-1}(L_A \| \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} + L_b), \quad and \quad \pi_p = L_{\nabla_\alpha S_u}.$$

Proof. By Lemma 4.1, we have

$$\zeta \| p^k - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} \geq \| p^{k+1} - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} - (L_A \| \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} + L_b) \gamma_N^{-1} \| u - S_u(\alpha) \|_Q.$$

Rearranging terms and multiplying by ζ^{-1} gives

$$\kappa_p \| p^{k+1} - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} \leq \| p^k - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} + C_S \| u - S_u(\alpha) \|_Q.$$

The triangle inequality and Lipschitz continuity of $\nabla_\alpha S_u$ gives

$$\| p^k - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}} \leq \| p^k - \nabla_\alpha S_u(\check{\alpha}) \|_{Q^{-1}} + \| \nabla_\alpha S_u(\check{\alpha}) - \nabla_\alpha S_u(\alpha) \|_{Q^{-1}}$$
$$\leq \| p^k - \nabla_\alpha S_u(\check{\alpha}) \|_{Q^{-1}} + L_{\nabla_\alpha S_u} \| \check{\alpha} - \alpha \|.$$

Combining the previous estimates finishes the proof. □

We needed to assume that the solution mapping of the inner problem is Lipschitz continuously differentiable to prove the adjoint tracking estimates. We discuss next the reasonability of this assumption. If the inner and adjoint solution mappings S_u and S_p are Lipschitz, then also $\nabla_\alpha S_u$ is since $\nabla_\alpha S_u(\alpha) = S_p(S_u(\alpha), \alpha)$. The Lipschitz properties of S_u are discussed in Remark 2.4. We next prove the Lipschitz continuity of S_p under similar assumption as in Lemma 4.1.

Lemma 4.10. *Let U and $\mathscr{A}$ be Hilbert spaces. Let further $Q \in \mathbb{L}(U; U)$ be positive definite, and $\nabla G : U \times \mathscr{A} \to \mathbb{L}(U; U \times \mathscr{A})$ be invertible and Lipschitz continuous with constant $L_{\nabla G}$ in some bounded closed set $V_u \times V_\alpha$. Moreover, assume that $\|(\nabla_u G(u; \alpha))^{-1}\| \le \gamma_G^{-1}$ in $V_u \times V_\alpha$ for some $\gamma_G > 0$. Then S_p defined in (2.3) is Lipschitz in $V_u \times V_\alpha$, i.e.*

$$\|S_p(u_1, \alpha_1) - S_p(u_2, \alpha_2)\|_{Q^{-1}} \le L_{S_p}(\|u_1 - u_2\|_Q + \|\alpha_1 - \alpha_2\|)$$

for $u_1, u_2 \in V_u$ and $\alpha_1, \alpha_2 \in V_\alpha$ with the factor

$$L_{S_p} := \gamma_G^{-1} L_{\nabla G}\left(1 + \gamma_G^{-1} \max_{(u,\alpha) \in V_u \times V_\alpha} \|\nabla_\alpha G(u, \alpha)\|_{Q^{-1}}\right).$$

Proof. Using the definition of S_p in (2.3), we rearrange

$$S_p(u_1, \alpha_1) - S_p(u_2, \alpha_2) = (\nabla_\alpha G(u_1; \alpha_1) - \nabla_\alpha G(u_2; \alpha_2))(\nabla_u G(u_1; \alpha_1))^{-1}$$
$$+ \nabla_\alpha G(u_2; \alpha_2)\left((\nabla_u G(u_1; \alpha_1))^{-1} - (\nabla_u G(u_2; \alpha_2))^{-1}\right).$$

Since Q is positive definite, hence invertible, we have $\|pu\|_{\mathscr{A}} = \|(pQ^{-1/2})(Q^{1/2}u)\| \le \|p\|_{Q^{-1}}\|u\|_Q$. This and the triangle inequality give

$$(4.11) \quad \|S_p(u_1, \alpha_1) - S_p(u_2, \alpha_2)\|_{Q^{-1}}$$
$$\le \|(\nabla_u G(u_1; \alpha_1))^{-1}\|\|\nabla_\alpha G(u_1; \alpha_1) - \nabla_\alpha G(u_2; \alpha_2)\|_{Q^{-1}}$$
$$+ \|\nabla_\alpha G(u_2; \alpha_2)\|_{Q^{-1}}\|(\nabla_u G(u_1; \alpha_1))^{-1} - (\nabla_u G(u_2; \alpha_2))^{-1}\| =: E_1 + E_2.$$

The bound $\|(\nabla_u G(u; \alpha))^{-1}\| \le \gamma_G^{-1}$ and the Lipschitz continuity of ∇G in $V_u \times V_\alpha$ give

$$(4.12) \quad E_1 \le \gamma_G^{-1} L_{\nabla_\alpha G}\left(\|u_1 - u_2\|_Q + \|\alpha_1 - \alpha_2\|\right).$$

Towards estimating the second term on the right hand side of (4.11), we observe that

$$A^{-1} - B^{-1} = A^{-1}BB^{-1} - A^{-1}AB^{-1} = A^{-1}(A - B)B^{-1}$$

for any invertible linear operators A, B. This and Lipschitz continuity of ∇G give

$$E_2 = \|\nabla_u G(u_1; \alpha_1)^{-1}(\nabla_u G(u_1; \alpha_1) - \nabla_u G(u_2; \alpha_2))\nabla_u G(u_2; \alpha_2)^{-1}\| \cdot \|\nabla_\alpha G(u_2; \alpha_2)\|_{Q^{-1}}$$
$$\le \left(\max_{(u,\alpha) \in V_u \times V_\alpha} \|\nabla_\alpha G(u, \alpha)\|_{Q^{-1}}\right)\|\nabla_u G(u_1; \alpha_1)^{-1}\|$$
$$\cdot \|\nabla_u G(u_2; \alpha_2)^{-1}\|\|(\nabla_u G(u_1; \alpha_1) - \nabla_u G(u_2; \alpha_2))\|$$
$$\le \gamma_G^{-2} L_{\nabla G}\left(\max_{(u,\alpha) \in V_u \times V_\alpha} \|\nabla_\alpha G(u, \alpha)\|_{Q^{-1}}\right)\left(\|u_1 - u_2\|_Q + \|\alpha_1 - \alpha_2\|\right).$$

Inserting this inequality and (4.12) into (4.11) establishes the claim since $\|\nabla_\alpha G(u, \alpha)\|_{Q^{-1}}$ is bounded in $V_u \times V_\alpha$ by the Lipschitz continuity of ∇G. $\qquad\qquad\square$

5 NUMERICAL EXPERIMENTS

In this section, we evaluate the operator identification performance of the following methods:

PDPS + block-GS Algorithm 5.1, i.e., our general bilevel approach with the PDPS of Example 3.2 for the inner problem, and the block Gauss–Seidel splitting of Example 4.8 for the adjoint.

PDPS + identity Our general bilevel approach with PDPS for the inner problem, and the identity splitting of Example 4.5 for the adjoint.

implicit Solve both the inner problem and the adjoint equation to a high precision.

trust region The inexact derivative-free method of [16] that solves inner problem inexactly while taking trust region steps to solve the bilevel (outer) problem.

In all cases except the trust region method, the outer iterates are updated with forward-backward steps.

Note that the algorithms PDPS + block-GS and PDPS + identity follow the framework of Algorithm 2.1, i.e. satisfy the inner and adjoint tracking conditions, if the objective functions satisfy the assumptions of Theorem 3.6 and the conditions discussed in the splitting Examples 4.5 and 4.7 respectively.

The implicit method is used as a baseline comparison and the trust region method as more advanced one. We have chosen the latter as a comparison from the more recent literature, as the algorithm itself requires the least information as a derivative-free method, and, therefore, can, in principle, be applied to the (primal-only form of) the same inner problem as our method, unlike many other methods that would require a different smoothing of the inner regularisation term g to achieve the required second-order differentiability, as we will discuss below. The specific convergence conditions of [16] can be difficult to verify, however, and might indirectly require the second-order differentiability of the inner regulariser.

Our specific operator identification problems are:

MRI Find an optimal sparse subsampling (line) pattern for MRI reconstruction using four slices of the digital brain phantom of [2] as a ground-truth in the outer problem, and corresponding simulated fully sampled MRI measurements as the data of the corresponding inner-problem(s).

deblurring Find an optimal parametrisation of convolution kernel and regularisation parameter to deblur/deconvolve a photograhic image with simulated blur.

We define component functions used in both problems in Section 5.1. Then in Sections 5.2 and 5.3 we present inner problem and adjoint equations details for MRI and deblurring respectively. We present numerical details in Section 5.4, and discuss performance in Section 5.5.

5.1 OBJECTIVE FUNCTIONS

For a training set of m "ground truth" images $\{b_i\}$ of dimension $n_1 \times n_2$, i.e. each $b_i \in \mathbb{R}_+^{n_1 n_2}$, we take as the outer data fitting term

$$J(u) = \frac{1}{2} \sum_{i=1}^{m} \|x_i - b_i\|_2^2, \qquad u_i = (x_i, y_i) \in \mathbb{R}_+^{n_1 n_2} \times \mathbb{R}^{2n_1 n_2}.$$

We specify the outer regulariser R and the (primal-dual) inner problem for the different experiments in corresponding subsections.

Algorithm 5.1 PDPS + block-GS

Require: Functions $J : U \to \mathbb{R}$, and $R : \mathscr{A} \to \overline{\mathbb{R}}$, with J Fréchet differentiable and R convex, on Hilbert spaces $U = X \times Y$ and $\mathscr{A}$. Functions $f_0, e : X \times \mathscr{A} \to \overline{\mathbb{R}}$ and $g : Y \times \mathscr{A} \to \overline{\mathbb{R}}$ convex in their first parameter, e has a L–Lipschitz gradient with respect to its first parameter. Also let $K \in \mathbb{L}(X; Y)$, Outer step length $\sigma > 0$, adjoint step length $\theta_y > 0$, inner step lengths $\tau_x, \tau_y > 0$ satisfying $\tau_x L/2 + \tau_x \tau_y \|K\|^2 \leq 1$ and parameter $\omega > 0$.

1: Pick an initial iterate $(u^0, p^0, \alpha^0) \in U \times \mathbb{L}(U; \mathscr{A}) \times \mathscr{A}$.

2: **for** $k \in \mathbb{N}$ **do**

3:      $x^{k+1} := \mathrm{prox}_{\tau_x f_0(\,\cdot\,;\alpha^k)}(x^k - \tau_x(K^* y^k + \nabla_x e(x^k; \alpha^k)))$ ▷ inner primal update

4:      $y^{k+1} := \mathrm{prox}_{\tau_y g^*(\,\cdot\,;\alpha^k)}(y^k + \tau_y K((1 + \omega)x^{k+1} - \omega x^k))$ ▷ inner dual update

5:      Choose a splitting $\nabla_x^2 f(x^{k+1}; \alpha^k) = N_{11} + M_{11}$ with N_{11} invertible.

6:      Write $N_{22} = \nabla_y^2 g^*(y^{k+1}; \alpha^k) + \theta_y^{-1} \mathrm{Id}$

7:      $p_x^{k+1} := N_{11}^{-1}\left(-M_{11}p_x^k - \nabla_\alpha \nabla_x f(x^{k+1}; \alpha^k) - K^* p_y^k\right)$ ▷ adjoint primal update

8:      $p_y^{k+1} := N_{22}^{-1}\left(\theta_y^{-1}p_y^k - \nabla_\alpha \nabla_y g^*(y^{k+1}; \alpha^k) - Kp_x^{k+1}\right)$ ▷ adjoint dual update

9:      $\alpha^{k+1} := \mathrm{prox}_{\sigma R}\left(\alpha^k - \sigma p^{k+1}\nabla_u J(u^{k+1})\right)$ ▷ outer forward-backward step

10: **end for**

For simplicity, we have not yet analysed our method in a stochastic setting, although a generalisation should be immediate.[3] We therefore concentrate on small sample sets, with the goal of evaluating the performance of our splitting approaches. These performance improvements should directly generalise to an eventual stochastic extension of our approach.

The general form of the inner problem, related primal-dual optimality conditions and description of the algorithm can be found in Example 3.2. We take $g(y; \alpha_0) = \alpha_0 \|y\|_{2,1}$, where $\|y\|_{2,1}$ is the sum over pixelwise two-norms. The use the steps of the PDPS for the inner problem, we need its Fenchel conjugate

$$g^*(y; \alpha_0) = \sum_{j=1}^{n_1 n_2} \delta_{B_{\mathbb{R}^2}(0,\alpha_0)}(y_j).$$

However, our theory does not yet allow for the inner problem to be nonsmooth.[4] Therefore, instead of $g^*(y; \alpha_0)$, we use the C^2 strongly convex approximation

$$(5.1) \qquad g_{\varepsilon,\delta}^*(y; \alpha_0) = \sum_{j=1}^{n_1 n_2} \max\{0, \frac{1}{3\varepsilon}(\|y_j\| - \alpha_0)^3\} + \frac{\delta}{2}\|y_j\|^2 \qquad (y_j \in \mathbb{R}^2)$$

Since $g_{\varepsilon,\delta}^*$ is convex, proper, and lower semicontinuous, it is the Fenchel conjugate of $g_{\varepsilon,\delta} := g_{\varepsilon,\delta}^{**}$. We take $\varepsilon = 1 \cdot 10^{-6}$ and $\delta = 1 \cdot 10^{-4}$, which are numerically practical small values that ensure that $g_{\varepsilon,\delta}^* \approx g^*$. We differentiate and derive the proximal operator of $g_{\varepsilon,\delta}^*$ in Appendix D.2

---

[3] This would simply involve selecting on each iteration k a random subset $S_k \in \{1, \ldots, m\}$, and making the step with respect to $\frac{m}{2\#S_k} \sum_{i \in S_k} \|x_i - b_i\|$ instead of J. Then steps for the inner problems would only have to be made for $i \in S_k$. However, as the inner and adjoint variables for $i \notin S_k$ would not be updated, convergence would need to be analysed carefully.

[4] Assumption 2.2 makes no such restriction, and we could, in principle, take S_u as a differentiable selection of a multi-valued solution map. However, to prove the assumption in Sections 3 and 4, we needed to impose differentiability and strong convexity assumptions.
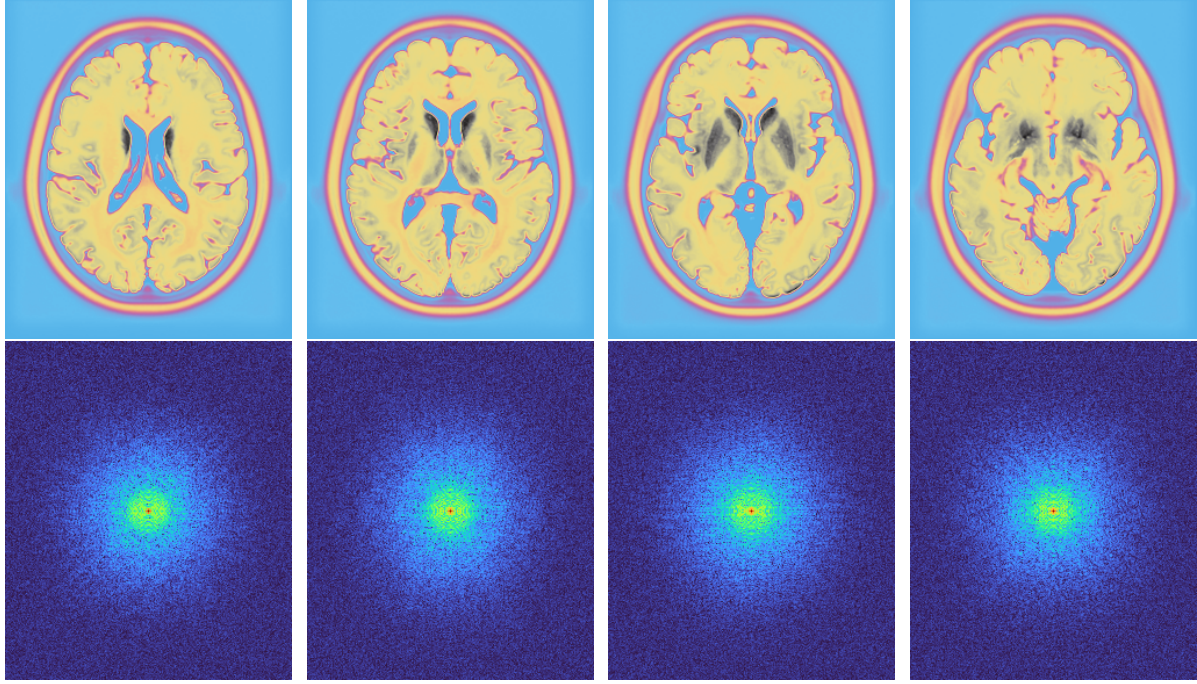

Figure 1: Top: true data (4 instances) presented in the colormap ▭ with values from interval $[0, 1]$. Bottom: corresponding simulated fully sampled MRI measurements presented in logarithmic scale, computed as $\log_{10}(|\cdot| + 0.05)$, using colormap ▭ with values from interval $[-1.3, 2.0]$.

5.2 MRI

The ground truth of the training set, $\{b_i\}$, contains four 247×292 slices of the digital brain phantom [2] scaled into interval $[0, 1]$ and corresponding simulated fully sampled MRI measurement $\{z_i\}$, see Figure 1. The measurements $\{z_i\}$ are simulated by adding Gaussian noise of standard deviation 0.02 to $\{b_i\}$ and then taking the Fourier transform.

The inner (inverse) problem is

$$(5.2) \qquad \arg\min_{x_i \in \mathbb{R}_+^{n_1 n_2}} \frac{1}{2} \|Z_\alpha(\mathcal{F}x_i - z_i)\|_2^2 + g_{\varepsilon,\delta}(Dx_i; \alpha_0) \qquad (\alpha \in [0,\infty)^{75})$$

where Z_α is the subsampling operator (diagonal matrix), see Figure 2a, $\mathcal{F}$ is discrete 2-D Fourier transform (matrix) and D is a backward difference operator (matrix) with Dirichlet boundary conditions. We take $\alpha_0 = 0.02$ constant because the subsampling operator Z_α can also model the regularisation factor, and we do not want several competing parameters to optimise. This value of α_0 was chosen by trial and error to achieve good reconstruction with fully sampled data, i.e. $Z_\alpha = \text{Id}$. To apply Example 3.2 for the derivation of the PDPS, we split $f = f_0 + e$ for $f_0(x; \alpha) = \frac{1}{2}\|Z_\alpha(\mathcal{F}x - z)\|_2^2$ and $e(x; \alpha) = 0$. Although f_0 involves the linear forward operator $Z_\alpha(\mathcal{F}$, its proximal mapping can be easily calculated due to the unitarity of $\mathcal{F}$ and the diagonality of Z_α.

Remark 5.1. This choice of f_0 is not necessarily strongly convex as required for Theorem 3.6 to ensure the inner tracking property. It is, however, convex, and due to our good numerical results, we have not added artificial strong convexity. Moreover, in many cases, appropriate local growth can be elicited through the metric subregularity of the entire objective [44, 43, 9].

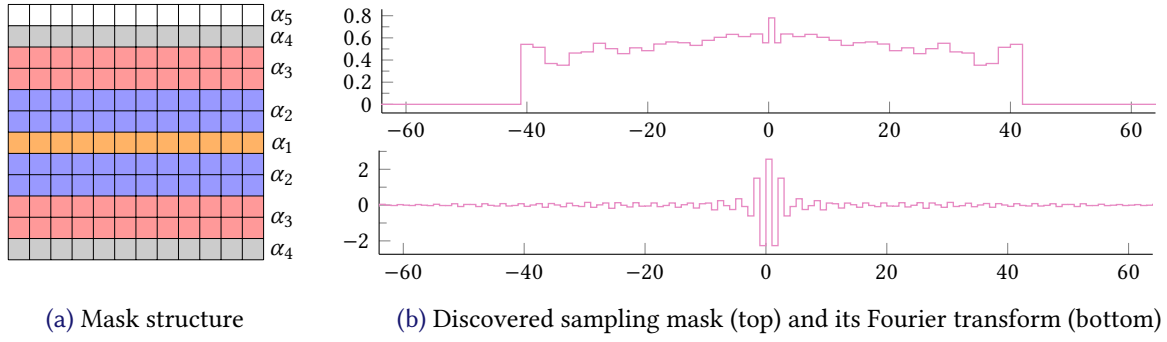We use block Gauss–Seidel splitting for the adjoint equation following Example 4.8. We have

(a) Mask structure

(b) Discovered sampling mask (top) and its Fourier transform (bottom)

Figure 2: Sampling pattern for MRI. Figure 2a shows how α parametrise the sampling pattern with a small example. Different colours represent different components of α such that applying the mask multiplies every pixel of one colour with same multiplier α_i. In our case image height is 292 and α has 75 elements. Figure 2b presents identified pattern weights and their Fourier transform, with the x-axis restricted to $[-64, 64]$.

$\nabla_x^2 f(u; \alpha) = \mathcal{F}^* Z_\alpha^2 \mathcal{F}$ and we take $N_{11} = \mathcal{F}^* Z_\theta^2 \mathcal{F}$ with $Z_\theta^2 = \max\{\theta_x^{-1}, Z_\alpha^2\}$ for

$$(5.3) \qquad \theta_x^{-1}(i, j) = 0.1 + 0.4 \left(1 - \sin\left(\tfrac{i}{n_1}\pi\right) \sin\left(\tfrac{j}{n_2}\pi\right)\right)^2, \quad (i \in \{1, ..., n_1\}, \; j \in \{1, ..., n_2\})$$

and the maximum is taken pointwise. To be more precise θ_x^{-1} is the matrix (5.3) stacked into the diagonal of $n_1 n_2 \times n_1 n_2$ matrix.

For the outer regulariser we take

$$R(\alpha) = \beta\varphi(w^T\alpha) + \delta_{[0,\infty)^n}(\alpha) \quad \text{where} \quad \varphi(t) = t + \delta_{(-\infty, M]}(t)$$

with the parameter $\beta = 10$ and sparsity control $M = 0.15$. The vector w, whose components sum to 1, weights the components of α according to how many lines of the k-space they correspond to, see Figure 2a. Our choice for the initial subsampling weights α^0 has each component 0.15 according to our choice for parameter M such that $\alpha^0 \in \mathrm{dom}\, R$ and α^0 doesn't contain prior information about important weights. The proximal operator of R we present in Appendix D.1.

5.3 DEBLURRING

We use a cropped portion (128×128) of image o2 from the free Kodak dataset [17] converted to gray values in $[0, 1]$ as our "ground truth" $\{b_i\}$. The 2-D convolution operator (matrix) parametrised by α_2, α_3 and α_4 as illustrated in Figure 3a is denoted by A_α. Operator r_θ rotates image θ degrees, clockwise for $\theta > 0$ and counterclocwise for $\theta < 0$. We form $\{z_i\}$ by computing $r_{-1}(A_\alpha r_1(b_i))$ for $[\alpha_2, \alpha_3, \alpha_4] = [0.15, 0.1, 0.75]$ and adding Gaussian noise of standard deviation 0.02.

The inner (inverse) problem is

$$(5.4) \qquad \arg\min_{x_i \in \mathbb{R}_+^{n_1 n_2}} \frac{1}{2}\|A_\alpha x_i - z_i\|_2^2 + g_{\varepsilon,\delta}(D x_i; \alpha_0) \qquad (\alpha \in [0, \infty)^4)$$

where A_α is the aforementioned convolution operator and D is a backward difference operator. We use the scaled regularisation parameter $\alpha_0 = C\alpha_1$ for the constant $C = \frac{1}{10}$ to help with outer iterate convergence by ensuring the same order of magnitude for all components of α. We consider f in Example 3.2 as sum of $f_0(x; \alpha) = 0$ and $e(x; \alpha) = \frac{1}{2}\|A_\alpha x_i - z_i\|_2^2$, which means that the primal step for the the inner problem does not involve a proximal operator, but only gradient step. Again this choice of f_0 is not necessarily strong convex; see, however, Remark 5.1.

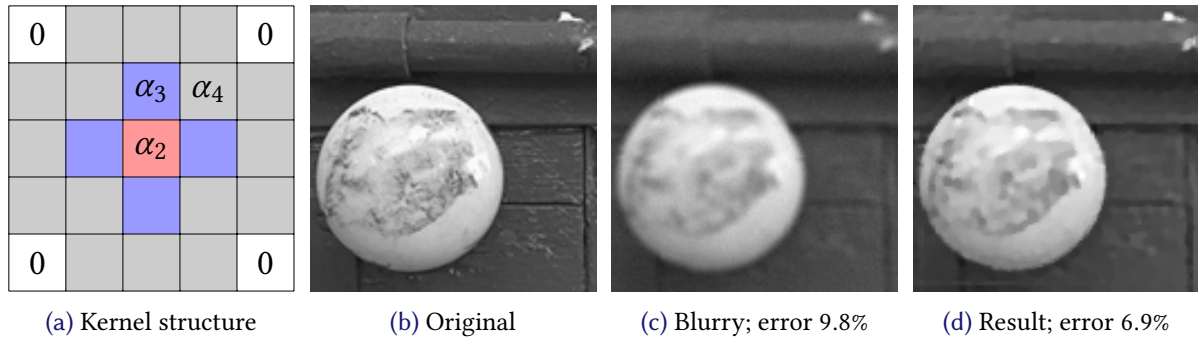(a) Kernel structure   (b) Original   (c) Blurry; error 9.8%   (d) Result; error 6.9%

Figure 3: Deconvolution kernel parametrisation, data, and result for PDPS + block-GS. The different colours in 3a represent different components of α such that elements of kernel with same colour have same value. The errors in 3c and 3d are $\|\text{image} - \text{target}\|_2 / \|\text{target}\|_2$.

For the block Gauss–Seidel splitting we need to choose an invertible N_{11}. We use the fact that $\nabla_x^2 f(u; \alpha) = A_\alpha^* A_\alpha$ equals $\mathcal{F}^* Z_\alpha^2 \mathcal{F}$ with Z_α being the Fourier transform of the convolution kernel, or more precisely it stacked into diagonal of $n_1 n_2 \times n_1 n_2$ matrix. This lets us again take $N_{11} = \mathcal{F}^* Z_\theta^2 \mathcal{F}$ such that $Z_\theta^2 = \max\{\theta_x^{-1}, Z_\alpha^2\}$ for θ_x^{-1} defined in (5.3).

With the parameter $\beta = 10^4$, our choice for the outer regulariser is

$$R(\alpha) = \beta \left(\alpha_2 + \alpha_3 + \alpha_4 - 1\right)^2 + \delta_{[0,\infty)}(\alpha_1).$$

The proximal operator of R we present in Appendix D.1.

5.4 NUMERICAL SETUP

Our algorithm implementations are provided on Zenodo [41]. The specific parameter choices for both experiments are listed in Table 1. To pick inner step lengths $\tau_x, \tau_y > 0$ satisfying $\tau_x L/2 + \tau_x \tau_y \|D\|^2 \leq 1$ for PDPS we use the upper bound of $\|D\|^2 \leq 8$ from [4]. All the other step lengths, numbers of step for solving inner problem and adjoint equation to a high precision in implicit method, and parameters, such as $\omega = 1$, are chosen by trial and error to obtain an apparently stable, but as efficient as possible, algorithm. We do not attempt to verify their theoretical conditions. The chosen initial outer iterates α^0 are specified in Sections 5.2 and 5.3. Initial inner and adjoint iterates for PDPS + block-GS and PDPS + identity are obtained by solving inner problem and adjoint equation to a high precision similarly than in implicit method for corresponding experiment.

The adjoint equation in the implicit method is solved with block Gauss–Seidel, see Example 4.7, for MRI and with the conjugate gradients squared method [39] for deblurring. For the latter, we use Matlab's cgs implementation with tolerance 10^{-4} and maximum iteration count 2000. The dimension of the adjoint equation in our MRI experiment is large, 64911600. Matlab's standard linear solvers, like cgs or bicgstab [45], did not scale well to such a high-dimensional problem, so we used block Gauss–Seidel instead. We took 200 steps of the latter, which still provides a less precise solution of the adjoint equation than cgs for deblurring.

The parameters of the trust region method in the deblurring experiment are also chosen by trial and error, as follows:

- initial and maximum trust region radii $\Delta^0 = 0.01$ and $\Delta_{\max} = 0.1$, as well as rates of change $\gamma_{\text{inc}} = 2$ and $\gamma_{\text{dec}} = 0.1$,
- step validation parameters $\eta_1 = 0.01$, $\eta_2 = 0.5$ and $\eta_1' = 0.004$,
- FISTA parameters for the inner problem, $\mu = 1 \cdot 10{-4}$, $L = 1 \cdot 10^5$, and $\tau = 1/L = 1 \cdot 10^{-5}$. On each outer step, the inner FISTA is initialized using a restarting strategyfrom the previous outer

Table 1: Algorithm parametrisation, time multiplier, and total outer steps to reach threshold computational resources (CPU time) value. The threshold is 15000 for MRI and 10000 for deblurring, except 6000 and 3000 for PDPS + block-GS. The time multipliers allow conversion from computational resources to seconds. It differs between algorithms and problems due to different levels of parallelisability. The 'inner steps' and 'adjoint steps' indicate the (maximum) number of iterations taken towards a solution of the inner problem or the adjoint equation on every outer iteration. The cgs method (of Matlab) that we use to solve the adjoint for the implicit method in the deblurring experiment, may use a smaller number of iterations as determined by the tolerance 10^{-4}. Similarly, the inner level solver of trust region method may take less iterations if it achieves the dynamic tolerance. Step lengths for PDPS (inner steps) are $\tau_x = 0.354, \tau_y = 0.350$ for MRI and $\tau_x = 0.600, \tau_y = 0.141$ for deblurring.

	method	outer steps	inner steps	adjoint steps	time mult.	θ_x, θ_y	σ
MRI	implicit	16	$3 \cdot 10^3$	200	0.23	(5.3), 0.1	$7 \cdot 10^{-4}$
	PDPS + identity	$1.9 \cdot 10^3$	1	1	0.23	$0.1, 6.25 \cdot 10^{-4}$	$1 \cdot 10^{-5}$
	PDPS + block-GS	$7.6 \cdot 10^2$	1	1	0.22	(5.3), 0.1	$1 \cdot 10^{-4}$
deblur	implicit	142	$2.5 \cdot 10^3$	$2 \cdot 10^3$	0.18	-	$2 \cdot 10^{-4}$
	PDPS + identity	$4.9 \cdot 10^2$	1	1	0.22	$1 \cdot 10^{-3}, 1 \cdot 10^{-3}$	$5 \cdot 10^{-7}$
	PDPS + block-GS	$5.2 \cdot 10^5$	1	1	0.21	(5.3), 0.1	$1 \cdot 10^{-5}$
	trust region	48	$2.5 \cdot 10^4$	-	0.18	-	-

step. It takes maximum of 25000 steps or until the tolerance $\|\text{inner gradient}\|^2 < 10^3 (\Delta^k)^2 \mu^2$ is achieved.

We do not present parameters and results for the trust region method for the MRI experiment, as it scaled poorly to such a large-dimensional problem. Already the first step of the method surpassed the limit 15000 of computational resources (CPU Time), common to all algorithms.

To compare algorithm performance, we plot relative errors and values of outer objective as function of the cputime value of Matlab on an AMD Ryzen 5 5600H CPU. We call this value "computational resources", since it takes measures the use of several CPU cores by Matlab's internal linear algebra. This is fairer than the elapsed real time.

We need estimates $\tilde{\alpha}$ and $\tilde{u}$ of optimal $\widehat{\alpha}$ and $\widehat{u} = S_u(\widehat{\alpha})$ to compare performance of the algorithms. For the MRI experiment these estimates are obtained by running PDPS + block-GS until computational resource (CPU time) value of 8000. For deblur experiment er use slightly different estimates for PDPS + block-GS and PDPS + identity, which are obtained by running the corresponding algorithms for CPU time values of 4000 and 15000. The first estimates are also used to track performance of implicit method. With these solution estimates we define the inner and outer relative errors

$$e_{\alpha,\text{rel}} := \frac{\|\tilde{\alpha} - \alpha^k\|}{\|\tilde{\alpha}\|} \quad \text{and} \quad e_{u,\text{rel}} := \frac{\|\tilde{u} - u^k\|_Q}{\|\tilde{u}\|_Q} \quad \text{for } Q := \begin{pmatrix} \tau_x^{-1} \text{Id} & -D^* \\ -D & \omega^{-1}\tau_y^{-1} \text{Id} \end{pmatrix}.$$

5.5 RESULTS

We report performance in Figure 5 and the image data and reconstructions in Figures 3 and 4. Figure 5 indicates that PDPS + block-GS has significantly lower computational costs than the other methods. It has the fastest convergence of the outer objective function as well as the inner and the outer iterates in both experiments, MRI and deblurring. PDPS + identity also seems to perform better than implicit function and trust region methods based on Figure 5, but its efficiency appears to be closer to them
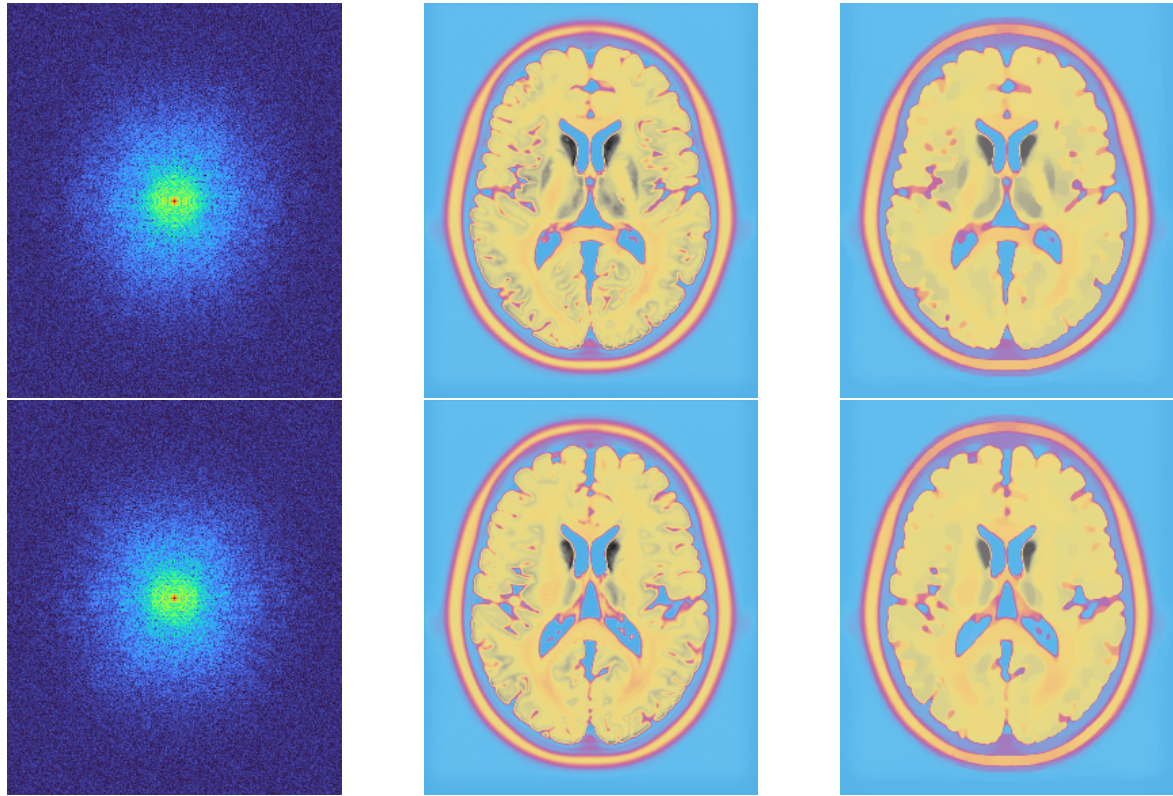
Figure 4: Target images of size 247×292 (middle column), simulated MRI measurements (left column) and reconstructions (right column). Colorbar for image domain ▬▬▬ values from interval $[0, 1]$ and for measurements presented in logarithmic scale computed as $\log_{10}(\cdot + 0.05)$ ▬▬▬ values (of logs) from interval $[-1.3, 2.0]$. Top row presents a training set example and bottom row test example, not used in training. The relative error $\|\text{reconstruction} - \text{target}\|_2 / \|\text{target}\|_2$ for the reconstruction of training target is 6.92% and for the test target 6.86%.

than to PDPS + block-GS. This suggests that block Gauss-Seidel for adjoint steps significantly improves algorithm performance. This improvement is likely based on fact that adjoint steps based on the block Gauss-Seidel and identity splitting have a similar computational cost, but the former allowed us to choose a 10–20 times larger outer step length parameter, see Table 1, while still obtaining an apparently stable algorithm.

All single-loop bilevel methods (as far as we know) require the inner objective to be twice continuously differentiable, including the methods we propose here. This means that nonsmooth inner problems need to be smoothed. This prevents us from (fairly) comparing our proposed method that uses PDPS steps for the inner problem, to most single-loop bilevel in the literature—including the FIFB and FEFB of [40]. Based on gradient steps for the inner problem, these would require a different type of smoothing. Regardless, in our previous work [40], for roughly the same deconvolution kernel identification problem that we treat here, we compared the performance of inner methods based on a single gradient descent step to the implicit method, based on the primal form of the inner problem. The implicit method present here, based on the primal-dual form of the inner problem, performs better than the implicit method in [40], but also, our single-loop method based on the block-GS adjoint solver, obtains greater improvements over the implicit method than the improvements obtained in [40].

From machine learning perspective, our MRI training set is extremely small, only 4 examples. To assess the quality of the identified sampling mask, we solved the inner inverse problem (5.2) to a high

precision with PDPS (3000 steps with the step length parameters from Table 1, since taking more steps did not decrease the inner objective value significantly) for simulated measurements z_i in a training set and in a test set of one data pair. Figure 4 shows these reconstructions as well as the true target images. These reconstructions lose target detail but are still of good quality (less than 7% relative error) for a sampling sparsity of 28%. The errors between the test set and the training set are similar, suggesting that our bilevel learning model generalises well to unseen data (from the same distribution) despite the small training set. We note that the reconstruction can be slightly improved by setting, as a post-processing step, all the non-zero learned weights to one.

As could be expected from our sparsity regularisation, for MRI, the reconstructed sampling mask in the k-space, shown in Figure 2b (top), takes a range of values. It emphasises the lowest frequencies, and then falls of until it reaches a threshold frequency, after which the mask is zero. The threshold frequencies appear to be again emphasised to compensate for the lack of sampling at higher frequencies. The Fourier transform of the mask, which would correspond to spatial regularisation of the image, resembles a sinc function, as would be expected from the structure of the k-space mask, but also, interestingly, the weights of a discretised higher-order differential operator. It, therefore, seems that an optimal MRI sampling mask attempts to extract information about the differentials of the image.

APPENDIX A THREE-POINT MONOTONICITY

We needed the following three-point monotonicity property in Lemma 2.9.

Theorem A.1. *Let X be Banach space and let $F : X \to \mathbb{R}$ be strongly convex with factor γ (in a set $A \subset X$) as well as Lipschitz differentiable with constant $L > 0$ (in A). Then for any $t > 0$,*

$$\langle DF(z) - DF(\widehat{x}), x - \widehat{x} \rangle \geq \frac{\gamma - tL}{2} \|x - \widehat{x}\|^2 + \frac{\gamma - tL}{2} \|z - \widehat{x}\|^2 - \frac{L}{4t} \|x - z\|^2 \quad (\text{for all } x, z, \widehat{x} \in A).$$

Proof. The proof is adapted from [42]. We have

$$
\begin{aligned}
(\text{A.1}) \qquad \langle DF(z) - DF(\widehat{x}), x - \widehat{x} \rangle &= \frac{1}{2} \langle DF(z) - DF(\widehat{x}), x - \widehat{x} \rangle + \frac{1}{2} \langle DF(z) - DF(\widehat{x}), x - \widehat{x} \rangle \\
&\geq \frac{1}{2} \Big(\langle DF(x) - DF(\widehat{x}), x - \widehat{x} \rangle + \langle DF(z) - DF(x), x - \widehat{x} \rangle \Big) \\
&\quad + \frac{1}{2} \Big(\langle DF(z) - DF(\widehat{x}), z - \widehat{x} \rangle + \langle DF(z) - DF(\widehat{x}), x - z \rangle \Big) \\
&=: \frac{1}{2} I_1 + \frac{1}{2} I_2.
\end{aligned}
$$

The strong monotonicity and Lipschitz continuity of DF with Young's inequality yield

$$
\begin{aligned}
I_1 &= \langle DF(x) - DF(\widehat{x}), x - \widehat{x} \rangle + \langle DF(z) - DF(x), x - \widehat{x} \rangle \\
&\geq \gamma \|x - \widehat{x}\|^2 - \|DF(z) - DF(x)\| \|x - \widehat{x}\| \geq \gamma \|x - \widehat{x}\|^2 - tL \|x - \widehat{x}\|^2 - \frac{L}{4t} \|x - z\|^2.
\end{aligned}
$$

Inserting this and an analogous estimate for I_2 into (A.1) establishes the claim. □

APPENDIX B LIPSCHITZ SELECTIONS OF THE SOLUTION MAP

We now continue from Remark 2.4 to treat the existence and differentiability of the solution map S_u of the inner problem. The next lemma is a an implicit function variant of [9, Lemma 22.3]. It relaxes the invertibility restriction on G_u in the standard implicit function theorem.
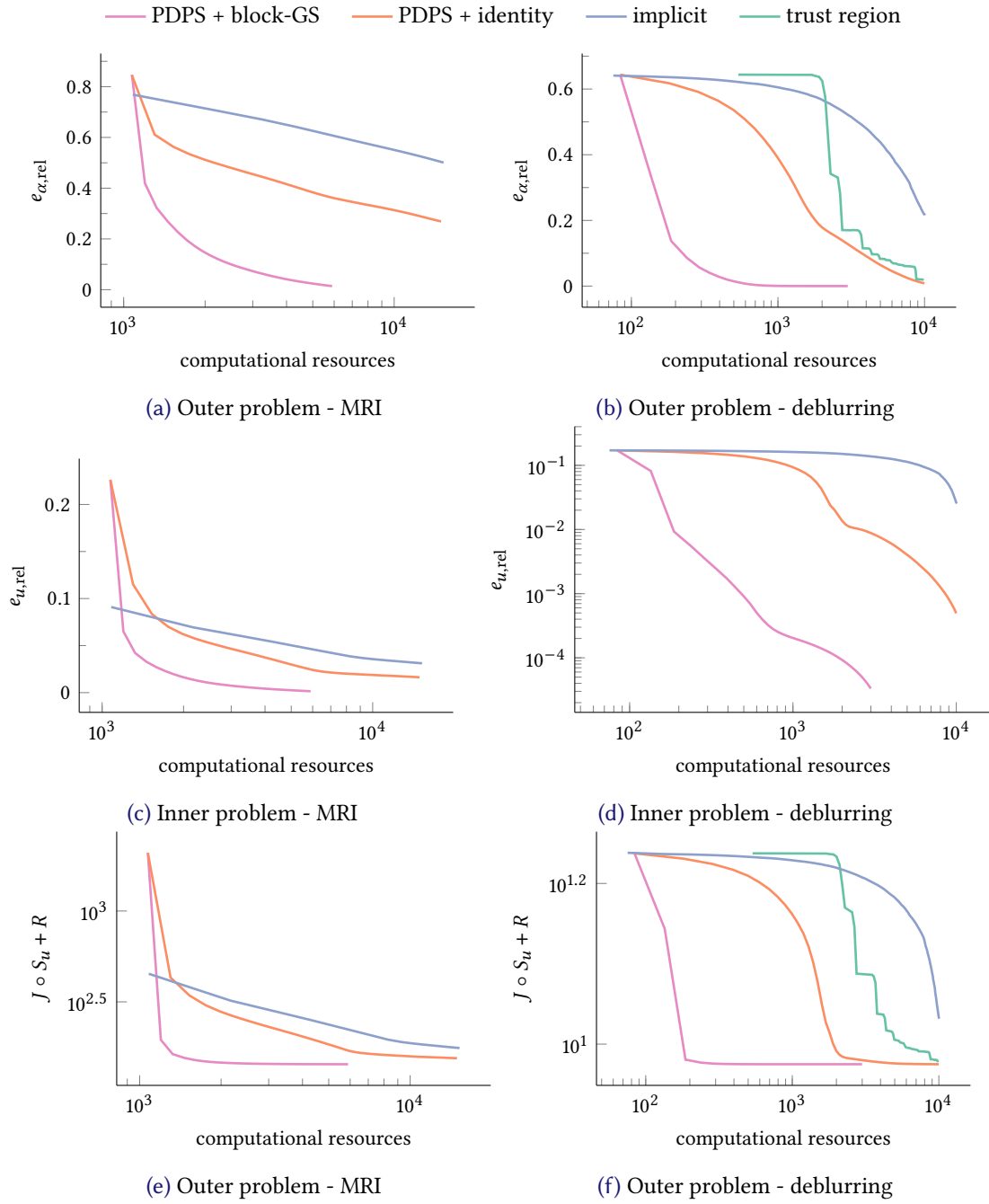
Figure 5: Performance of the compared methods. Top row: outer iterate convergence, middle row: inner iterate convergence and bottom row: outer function value convergence. The "computational resources" is the spent CPU time over multiple cores, parallelisation level depending on algorithm.

Lemma B.1. *On Banach spaces U, $\mathscr{A}$, and Y, suppose $G : U \times \mathscr{A} \to Y$ is continuously differentiable at (u, α), $G(u; \alpha) = 0$, and that $G_u(u; \alpha) \in \mathbb{L}(U; Y)$ has a right-inverse $G_u(u; \alpha)^\dagger \in \mathbb{L}(Y; U)$. Then there exists a neigborhood V_α of α and a continuously differentiable $S : V_\alpha \to U$ such that $G(S(\tilde{\alpha}); \tilde{\alpha}) = 0$ for all $\tilde{\alpha} \in V_\alpha$, and $S'(\alpha) = G_u(u; \alpha)^\dagger G_\alpha(u; \alpha)$.*

Proof. Let $A := G_u(u; \alpha)$ and $A^\dagger := G_u(u; \alpha)^\dagger$. Then $P := \mathrm{Id} - A^\dagger A$ is a projection into $\ker A =$

$\ker G_u(u; \alpha)$, in particular, $AP = 0$. We further define

$$\bar{G} : U \times \mathscr{A} \to Y \times \ker G_u(u; \alpha), \qquad \bar{G}(\tilde{u}; \tilde{\alpha}) := (G(\tilde{u}; \tilde{\alpha}), P\tilde{u}) \quad \text{for all } (\tilde{u}, \tilde{\alpha}) \in U \times \mathscr{A},$$

as well as

$$M : Y \times \ker A \to U, \qquad M(\tilde{y}, \tilde{q}) := A^\dagger \tilde{y} + \tilde{q}, \quad \text{for all } \tilde{y} \in Y \text{ and } \tilde{q} \in \ker A.$$

Then for all $\Delta u \in U$, we have $M\bar{G}_u(u; \alpha)\Delta u = A^\dagger A \Delta u + P\Delta u = \Delta u$. Thus M is a left-inverse of $\bar{G}_u(u; \alpha)$, and consequently $\ker \bar{G}_u(u; \alpha) = \{0\}$. Since $\bar{G}_u(u; \alpha)\Delta u = (A\Delta u, P\Delta u)$ for all $\Delta u \in U$, similarly, for all $(\tilde{y}, \tilde{q}) \in Y \times \ker A$, we have

$$\bar{G}_u(u; \alpha)M(\tilde{y}, \tilde{q}) = (AA^\dagger \tilde{y} + A\tilde{q}, PA^\dagger \tilde{y} + P\tilde{q}) = (AA^\dagger \tilde{y}, P\tilde{q}) = (\tilde{y}, \tilde{q}),$$

which shows that M is also the right-inverse of $\bar{G}_u(u; \alpha)$ on $Y \times \ker F'(x)$. Hence $\bar{G}_u(u; \alpha)$ is bijective. By the implicit function theorem, e.g., [27, Theorem 1.25], a continuously differentiable $S : V_\alpha \to U$ now exists in a neighborhood V_α of α with

$$\begin{aligned} S'(\alpha) &= \bar{G}_u(u; \alpha)^{-1}\bar{G}_\alpha(u; \alpha) = M\bar{G}_\alpha(u; \alpha) = M(G_\alpha(u; \alpha), 0) \\ &= A^\dagger G_\alpha(u; \alpha) = G_u(u; \alpha)^\dagger G_\alpha(u; \alpha). \end{aligned} \qquad \square$$

We now obtain the following; compare [14, Theorem 5J.8].

Corollary B.2. *Suppose that the assumptions of Lemma B.1 hold at some u_α with $G(u_\alpha; \alpha) = 0$ for all α in a compact $\check{\mathscr{A}} \subset \mathscr{A}$. If, moreover*
 (i) The set $\{u \mid G(u; \alpha) = 0\}$ is convex for all $\alpha \in \check{\mathscr{A}}$, and
 (ii) $M := \sup_{\alpha \in \check{\mathscr{A}}} \|G_u(u_\alpha; \alpha)^\dagger G_\alpha(u_\alpha; \alpha)\| < \infty$,
then for every $\varepsilon > 0$, there exists a continuously differentiable and Lipschitz $S : \check{\mathscr{A}} \to U$ with $G(S(\alpha); \alpha) = 0$ for all $\alpha \in \check{\mathscr{A}}$. Moreover, the Lipschitz factor of S is at most $M + \varepsilon$.

Proof. We apply Lemma B.1 to all $\alpha \in \check{\mathscr{A}}$ to obtain corresponding neighbourhoods $V_\alpha \subset \check{\mathscr{A}}$ and continuously differentiable mappings $S^\alpha : V_\alpha \to U$ satisfying $G(S^\alpha(\tilde{\alpha}); \tilde{\alpha}) = 0$ for all $\tilde{\alpha} \in V_\alpha$. Due to continuous differentiability and the expression $(S^\alpha)'(\alpha) = G_u(u; \alpha)^\dagger G_\alpha(u; \alpha)$, shrinking V_α if necessary, we may assume that S^α is Lipschitz in V_α with Lipschitz factor at most $\|G_u(u; \alpha)^\dagger G_\alpha(u; \alpha)\| + \varepsilon$.

If $\tilde{\alpha} \in V_{\alpha_1} \cap V_{\alpha_2}$ for $\alpha_1 \neq \alpha_2$, it may be that $S^{\alpha_1}(\tilde{\alpha}) \neq S^{\alpha_2}(\tilde{\alpha})$. However, the convexity assumption (i) guarantees for any $\lambda \in [0, 1]$ that $u^\lambda := \lambda S^{\alpha_1}(\tilde{\alpha}) + (1 - \lambda)S^{\alpha_2}(\tilde{\alpha})$ satisfies $G(u^\lambda; \tilde{\alpha}) = 0$. Therefore, we may use a compact covering argument and a partition of unity (which exists by the compactness of $\check{\mathscr{A}}$) to glue together the various S^α to obtain a single inverse selection $S : \check{\mathscr{A}} \to U$, satisfying $G(S(\alpha); \alpha) = 0$ for all $\tilde{\alpha}$. Finally, (ii) establishes the claimed bound on the Lipschitz factor. $\square$

If $G(u; \alpha) = \nabla_u F(u; \alpha)$ for a convex function F, then it is clear that the convexity assumption in Corollary B.2 (i) holds. However, it also holds for $G(u; \alpha) = A(u; \alpha) + B(u; \alpha)$ as given by Example 3.2 for the PDPS. As a variation of [22, Theorem 2.1], this can be seen from the characterisation of solutions $(\bar{x}, \bar{y})$ to $0 = G(\bar{x}, \bar{y}; \alpha)$ as the saddle points of $\mathcal{L}(x, y) := f(x; \alpha) + \langle Kx, y \rangle - g^*(y; \alpha)$. The latter as defined as those $(\bar{x}, \bar{y})$ satisfying

$$\mathcal{L}(\bar{x}, y) \leq \mathcal{L}(\bar{x}, \bar{y}) \leq \mathcal{L}(x, \bar{y}) \quad \text{for all} \quad (x, y) \in X \times Y.$$

Lemma B.3. *The saddle points of a convex-concave $\mathcal{L} : X \times Y \to [-\infty, \infty]$ form a convex set.*

Proof. For any two solutions pairs (x_0, y_0) and (x_1, y_1), defining $x_\lambda := \lambda x_0 + (1 - \lambda)x_1$ and $y_\lambda := \lambda y_0 + (1 - \lambda)y_1$, by convexity $\mathcal{L}(x_\lambda, y_\lambda) \leq \lambda \mathcal{L}(x_0, y_\lambda) + (1 - \lambda)\mathcal{L}(x_1, y_\lambda) \leq \mathcal{L}(x, y_\lambda)$ for any x. The lower bound is proved analogously. $\square$

APPENDIX C LEMMAS FOR THE TRACKING PROPERTY OF BLOCK GAUSS–SEIDEL

We prove here an estimate that was needed to derive the tracking inequality for block Gauss–Seidel splitting in Example 4.7. We start with two technical lemmas.

Lemma c.1. *Let $b > 0$ and $c \leq 0$. Then $-b + \sqrt{b^2 - 4bc} \geq -2bc/(-c + b)$.*

Proof. We estimate

$$-b + \sqrt{b^2 - 4bc} = \frac{b^2 - 4bc - b^2}{b + \sqrt{b^2 - 4bc}} = \frac{-4bc}{b + \sqrt{b^2 - 4bc}} \geq \frac{-4bc}{b + \sqrt{b^2 - 4bc + (2c)^2}} = \frac{-2bc}{-c + b}. \qquad \square$$

Lemma c.2. *Let $a, b, c, d > 0$ and define for $\eta_1, \eta_2 > 0$ the functions*

$$f_1(\eta_1, \eta_2) = (1 + \eta_1^{-1} + (1 + \eta_2^{-1})a^2)b^2 \quad and \quad f_2(\eta_1, \eta_2) = (1 + \eta_1)c^2 + (1 + \eta_2)d^2.$$

Moreover, assume for some $\zeta \in [0, 1)$ that

$$(\text{c.1}) \qquad\qquad d^2 + abd + b^2(1 + c)(1 + a^2) + c^2 + bc \leq \zeta^2$$

Then there exists $\eta_1^* > 0$ such that

$$f_1(\eta_1^*, acd^{-1}\eta_1^*) = f_2(\eta_1^*, acd^{-1}\eta_1^*) \leq \zeta^2.$$

Proof. The equation $f_1(\eta_1^*, acd^{-1}\eta_1^*) = f_2(\eta_1^*, acd^{-1}\eta_1^*)$ is equivalent to the quadratic

$$g(c^2 + acd)(\eta_1^*)^2 + (c^2 + d^2 - b^2(1 + a^2))\eta_1^* - b^2/c^2(c^2 + acd) = 0.$$

The left-hand-side is negative at $\eta_1^* = 0$, and the equation has a root $\eta_1^* > 0$. Instead of solving for this root, we find the root greater than η_1^* of

$$q(\eta) = (c^2 + acd)\eta^2 - b^2(1 + a^2)\eta - b^2/c^2(c^2 + acd).$$

This is

$$\eta = \frac{b^2(1 + a^2) + \sqrt{(b^2(1 + a^2))^2 + 4b^2/c^2(c^2 + acd)^2}}{2(c^2 + acd)}$$

$$\leq \frac{b^2(1 + a^2) + b/c(c^2 + acd)}{c^2 + acd} = \frac{b}{c}r \quad for \quad r := \left(1 + \frac{b(1 + a^2)}{c + ad}\right).$$

Because $f_2(\eta_1, (ac/d)\eta_1)$ is an increasing function of η_1, an application of (c.1) gives

$$f_2(\eta_1^*, (ac/d)\eta_1^*) \leq f_2(\eta, (ac/d)\eta) = \left(1 + \frac{b}{c}r\right)c^2 + \left(1 + \frac{ab}{d}r\right)d^2$$

$$\leq d^2 + abd + b^2(1 + c)(1 + a^2) + c^2 + bc \leq \zeta^2. \qquad \square$$

Theorem c.3. *Assume A_{11}, A_{12}, A_{21} and A_{22} are linear operator between Hilbert spaces U and $\mathcal{A}$. With $A_{11} = N_{11} + M_{11}$ and $A_{22} = N_{22} + M_{22}$, let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad N = \begin{pmatrix} N_{11} & 0 \\ A_{21} & N_{22} \end{pmatrix}, \quad and \quad M = \begin{pmatrix} M_{11} & A_{12} \\ 0 & M_{22} \end{pmatrix}.$$

Moreover, suppose N_{11} and N_{22} are invertible and bounded, and

$$(\text{c.2}) \quad \|N_{22}^{-1}(M_{22} - A_{21}N_{11}^{-1}A_{12})\|^2 + \|N_{22}^{-1}(M_{22} - A_{21}N_{11}^{-1}A_{12})\|\|N_{22}^{-1}A_{21}\|\|N_{11}^{-1}M_{11}\|$$

$$+ \|N_{11}^{-1}M_{11}\|^2(1 + \|N_{11}^{-1}A_{12}\|)(1 + \|N_{22}^{-1}A_{21}\|^2) + \|N_{11}^{-1}A_{12}\|^2 + \|N_{11}^{-1}A_{12}\|\|N_{11}^{-1}M_{11}\| \leq \zeta^2$$

for some $\zeta \in [0, 1)$. Then (4.7) holds with the aforementioned ζ and any γ_N satisfying

$$0 \leq \gamma_N \leq \frac{\|N_{11}\|\|N_{22}\|}{2\|N_{11}\| + \|N_{22}\|(1 + \|N_{22}^{-1}A_{21}\|^2)}.$$

Proof. We have

$$N^{-1} = \begin{pmatrix} N_{11}^{-1} & 0 \\ -N_{22}^{-1} A_{21} N_{11}^{-1} & N_{22}^{-1} \end{pmatrix}.$$

Thus Young's inequality, for any $\beta > 0$, establishes

(c.3) $$\|N^{-1}\|^2 = \sup_{x,y} \frac{\|N_{11}^{-1}x\|^2 + \|N_{22}^{-1}y - N_{22}^{-1}A_{21}N_{11}^{-1}x\|^2}{\|x\|^2 + \|y\|^2} = \sup_{x,y} \frac{\|x\|^2 + \|y - N_{22}^{-1}A_{21}x\|^2}{\|N_{11}x\|^2 + \|N_{22}y\|^2}$$

$$\leq \sup_{x,y} \frac{\|x\|^2 + (1+\beta^{-1})\|y\|^2 + (1+\beta)\|N_{22}^{-1}A_{21}x\|^2}{\|N_{11}x\|^2 + \|N_{22}y\|^2}.$$

Consequently $\rho\|N^{-1}\|^2 \leq 1$ (here $\rho = \gamma_N^2$) if and only if, for all x and y,

$$\|x\|^2 + (1+\beta)\|N_{22}^{-1}A_{21}x\|^2 \leq \rho^{-1}\|N_{11}x\|^2 \quad \text{and} \quad (1+\beta^{-1})\|y\|^2 \leq \rho^{-1}\|N_{22}y\|^2.$$

This is to say

(c.4) $$\rho + \rho(1+\beta)\|N_{22}^{-1}A_{21}\|^2 \leq \|N_{11}\|^2 \quad \text{and} \quad \rho(1+\beta^{-1}) \leq \|N_{22}\|^2.$$

We can solve from the latter

$$\beta^{-1} = \rho^{-1}\|N_{22}\|^2 - 1 \iff \beta = \frac{\rho}{\|N_{22}\|^2 - \rho} \iff \rho = \frac{\|N_{22}\|^2}{1+\beta^{-1}} = \frac{\beta\|N_{22}\|^2}{\beta+1}$$

provided $\|N_{22}\|^2 > \rho$. Then the first inequality in (c.4) reads

$$\frac{\beta\|N_{22}\|^2}{\beta+1} + \frac{\beta\|N_{22}\|^2}{\beta+1}(1+\beta)\|N_{22}^{-1}A_{21}\|^2 \leq \|N_{11}\|^2.$$

In other words

$$\beta\|N_{22}\|^2 + (1+\beta)\beta\|N_{22}\|^2\|N_{22}^{-1}A_{21}\|^2 \leq (1+\beta)\|N_{11}\|^2.$$

Note that this holds with $\beta = 0$, and rearranges as

(c.5) $$\beta^2\|N_{22}\|^2\|N_{22}^{-1}A_{21}\|^2 + \beta\left(\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2) - \|N_{11}\|^2\right) - \|N_{11}\|^2 \leq 0.$$

We want maximal ρ, therefore minimal $\beta > 0$ that satisfies (c.5). Since this is condition is quadratic in β, and is satisfied at $\beta = 0$,, the corresponding function has a positive root β'. Any positive number smaller than β', such as the positive root of

$$\beta^2\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2) + \beta\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2) - \|N_{11}\|^2,$$

also satisfies (c.5). We solve the corresponding quadratic equation in β and use Lemma c.1 to find $\beta^* < \beta'$

$$\frac{-\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2) + \sqrt{\|N_{22}\|^4(1+\|N_{22}^{-1}A_{21}\|^2)^2 - 4\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2)\|N_{11}\|^2}}{2\|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2)}$$

$$\geq \frac{\|N_{11}\|^2}{\|N_{11}\|^2 + \|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2)} =: \beta^*.$$

Therefore $\rho\|N_v^{-1}\|^2 \leq 1$ for $\rho = \beta^*\|N_{22}\|^2/(1+\beta^*)$ and we get

$$\rho = \frac{\beta^*\|N_{22}\|^2}{\beta^*+1} = \frac{\|N_{11}\|^2\|N_{22}\|^2}{2\|N_{11}\|^2 + \|N_{22}\|^2(1+\|N_{22}^{-1}A_{21}\|^2)} \geq \frac{\|N_{11}\|^2\|N_{22}\|^2}{\left(2\|N_{11}\| + \|N_{22}\|(1+\|N_{22}^{-1}A_{21}\|^2)\right)^2}$$

Taking the square root, we see that $\gamma_N \|N^{-1}\| \le 1$ for any

$$\gamma_N \le \|N_{11}\|\|N_{22}\| / (2\|N_{11}\| + \|N_{22}\|(1 + \|N_{22}^{-1}A_{21}\|^2)).$$

Likewise we have

$$N_v^{-1}M_v = \begin{pmatrix} N_{11}^{-1}M_{11} & N_{11}^{-1}A_{12} \\ -N_{22}^{-1}A_{21}N_{11}^{-1}M_{11} & -N_{22}^{-1}A_{21}N_{11}^{-1}A_{12} + N_{22}^{-1}M_{22} \end{pmatrix}.$$

Similarly as in (c.3), using Young's inequality we obtain $\|N_v^{-1}M_v\|^2 \le \zeta^2$ if and only if, for some $\eta_1, \eta_2 > 0$ and $\zeta \in [0, 1)$, we have

$$\left(1 + \eta_1^{-1} + (1 + \eta_2^{-1})\|N_{22}^{-1}A_{21}\|^2\right)\|N_{11}^{-1}M_{11}\|^2 \le \zeta^2 \quad \text{and}$$
$$(1 + \eta_1)\|N_{11}^{-1}A_{12}\|^2 + (1 + \eta_2)\|N_{22}^{-1}(M_{22} - A_{21}N_{11}^{-1}A_{12})\|^2 \le \zeta^2.$$

By Lemma c.2 these inequalities hold when the assumed (c.2) does. $\qquad\square$

APPENDIX D PROXIMAL OPERATORS

APPENDIX D.1 OUTER PROXIMAL OPERATORS

Lemma D.1. *Let $R(\alpha) = \beta(\sum_{i=2}^4 \alpha_i - 1)^2 + \delta_{[0,\infty)}(\alpha_1)$ for $\alpha = (\alpha_1, \dots, \alpha_4) \in \mathbb{R}^4$. Then*

$$\mathrm{prox}_{\sigma R}(\tilde{a}) = (\max\{0, \tilde{\alpha}_1\}, \alpha_2, \alpha_3, \alpha_4)$$

where $\alpha_2, \alpha_3,$ and α_4 are solved from the system of linear equations

$$\alpha_j + 2\sigma\beta \sum_{i=2}^4 \alpha_j = \tilde{a}_j + 2\sigma\beta.$$

Proof. The claim follows directly from the definition of proximal operator and the corresponding first order optimality condition

$$\partial_{\alpha_j}\left[\sigma\beta(\sum_{i=2}^4 \alpha_i - 1)^2 + \frac{1}{2}\sum_{i=2}^4(\alpha_i - \tilde{\alpha}_i)^2\right] = 0 \quad \text{for } j = 2, 3, 4. \qquad\square$$

The next result on our sparsity regulariser for MRI sampling pattern weights has a lot of resemblance to simplex projection methods, see, for example, [1, Chapter 5].

Lemma D.2. *For some weight vector $w \in (0, \infty)^n$ and $\beta, M > 0$, let*

$$R(\alpha) := \beta\varphi(w^\top\alpha) + \delta_{[0,\infty)^n}(\alpha), \quad \varphi(t) = t + \delta_{(-\infty, M]}(t).$$

Then, for all $i = 1, \dots, n$ and $\tilde{\alpha} \in \mathbb{R}^n$,

$$[\mathrm{prox}_{\tau R}(\tilde{\alpha})]_i = \max(0, \tilde{\alpha}_i - w_i\tilde{\lambda}) \quad \text{where} \quad \tilde{\lambda} = \max\left\{\frac{w_{\mathcal{I}}^\top\tilde{\alpha}_{\mathcal{I}} - M}{\|w_{\mathcal{I}}\|^2}, \tau\beta\right\},$$

where $\tau > 0$, $\tilde{\alpha}_{\mathcal{I}} = (\tilde{\alpha}_i)_{i \in \mathcal{I}}$ and $w_{\mathcal{I}} = (w_i)_{i \in \mathcal{I}}$ for $\mathcal{I} \subset \{1, \dots, n\}$ chosen such that

$$(\text{D.1}) \qquad\qquad \tilde{\alpha}_i/w_i > \tilde{\lambda} \implies i \in \mathcal{I} \quad \text{and} \quad \tilde{\alpha}_i/w_i < \tilde{\lambda} \implies i \notin \mathcal{I}.$$

Moreover, such a choice exists.

Thus, due to (D.1), to calculate α, we need to sort the vector $(\alpha_i/w_i)_{i=1}^n$, and find a dividing index j such that $\mathcal{I} = \{j+1, \ldots, n\}$ and $\tilde{\lambda}$ computed based on it satisfies (D.1). If $\alpha_i/w_i = \alpha_j/w_j$ and $w_i \neq w_j$ for $i \neq j$, then all sorting combinations may be needed to be tried for such indices. If $w^\top \alpha < M$, we can always take $\mathcal{I} = \{1, \ldots, n\}$.

Proof. The necessary and sufficient first-order optimality condition for $\alpha = \operatorname{prox}_{\tau R}(\tilde{\alpha})$ is

$$0 \in \alpha - \tilde{\alpha} + w N_{\leq M}(w^\top \alpha) + \tau \beta w + N_{\geq 0}(\alpha),$$

in other words

$$(D.2) \qquad 0 = \alpha - \tilde{\alpha} + w(\lambda + \tau \beta) + \mu \quad \text{where}$$

$$(D.3) \qquad 0 \leq \lambda, \quad w^\top \alpha \leq M, \quad (M - w^\top \alpha)\lambda = 0, \quad \text{and}$$

$$(D.4) \qquad \mu_i \leq 0, \quad \alpha_i \geq 0, \quad \alpha_i \mu_i = 0 \quad \text{for all} \quad i = 1, \ldots, n.$$

We choose

$$\lambda = \tilde{\lambda} - \tau \beta = \max \left\{ \frac{w_{\mathcal{I}}^\top \tilde{\alpha}_{\mathcal{I}} - M}{\|w_{\mathcal{I}}\|^2} - \tau \beta, 0 \right\}.$$

For any $i \in \{1, \ldots, n\}$, if $\tilde{\alpha}_i \geq w_i \tilde{\lambda}$, then $\alpha_i = \tilde{\alpha}_i - w_i \tilde{\lambda} \geq 0$, so (D.4) is satisfied by taking $\mu_i = 0$, and (D.2) when $\alpha_i - \tilde{\alpha}_i + w_i \lambda + \tau \beta w_i = 0$, which indeed holds. Likewise, if $\tilde{\alpha}_i < w_i \tilde{\lambda}$, we have $\alpha_i = 0$, so (D.4) is satisfied for any $\mu_i \leq 0$, and, therefore, (D.2) when $\alpha_i - \tilde{\alpha}_i + w_i \lambda + \tau \beta w_i \geq 0$, which again holds.

It, therefore, remains to verify (D.3). To do so, we start by proving that

$$(D.5) \qquad w^\top \alpha = w_{\mathcal{I}}^\top \tilde{\alpha}_{\mathcal{I}} - \|w_{\mathcal{I}}\|^2 \tilde{\lambda},$$

Indeed, by the construction $\alpha_i = \max(0, \tilde{\alpha}_i - w_i \tilde{\lambda})$, we have

$$w^\top \alpha = \sum_{i: \tilde{\alpha}_i \geq w_i \tilde{\lambda}} w_i \tilde{\alpha}_i - w_i^2 \tilde{\lambda} = \sum_{i: \tilde{\alpha}_i > w_i \tilde{\lambda}} w_i \tilde{\alpha}_i - w_i^2 \tilde{\lambda}.$$

Therefore (D.5) holds by (D.1). Now, to prove (D.3), observe that $0 \leq \lambda$ by construction. Using (D.5), and the definition of $\tilde{\lambda}$, we obtain $w^\top \alpha = \min(M, w_{\mathcal{I}}^\top \tilde{\alpha}_{\mathcal{I}} - \|w_{\mathcal{I}}\|^2 \tau \beta)$. In particular $w^\top \alpha \leq M$. Moreover, if $\lambda > 0$, i.e., $w_{\mathcal{I}}^\top \tilde{\alpha}_{\mathcal{I}} - \|w_{\mathcal{I}}\|^2 \tau \beta > M$, we obtain $w^\top \alpha = M$, proving (D.3).

To see that (D.1) can be satisfied, let $\mathcal{J} := \{i \mid \mu_i = 0\}$. Observe from (D.4) that $\mu_i = 0$ whenever $\alpha_i > 0$. Since $w_{\mathcal{J}}^\top \alpha_{\mathcal{J}} = w^\top \alpha \leq M$ and $w_{\mathcal{J}}^\top \mu_{\mathcal{J}}$, applying $w_{\mathcal{J}}^\top$ to the components of (D.2) corresponding to $\mathcal{J}$, we obtain

$$(D.6) \qquad 0 \leq M - w_{\mathcal{J}}^\top \tilde{\alpha}_{\mathcal{J}} + \|w_{\mathcal{J}}\|^2 (\lambda + \tau \beta).$$

Due to this and the first part of (D.3),

$$\lambda \geq \max \left\{ 0, \frac{w_{\mathcal{J}}^\top \tilde{\alpha}_{\mathcal{J}} - M}{\|w_{\mathcal{J}}\|^2} - \tau \beta \right\}.$$

If $w^\top \alpha < M$, we need $\lambda = 0$, so this must hold as an equality. If $w^\top \alpha = M$, the inequality in (D.6) would be an equality, so we again reach the same conclusion due to the non-negativity requirement. It follows from (D.2) and (D.4) that

$$\alpha_i = \tilde{\alpha}_i - w_i(\lambda + \tau \beta) - \mu_i = \tilde{\alpha}_i - w_i \tilde{\lambda} - \mu_i = \max\{0, \tilde{\alpha}_i - w_i \tilde{\lambda}_{\mathcal{J}}\},$$

where

$$\tilde{\lambda}_{\mathcal{J}} = \lambda + \tau \beta = \max \left\{ \frac{w_{\mathcal{J}}^\top \tilde{\alpha}_{\mathcal{J}} - M}{\|w_{\mathcal{J}}\|^2}, \tau \beta \right\}.$$

Finally, if $\tilde{\alpha}_i > w_i \tilde{\lambda}_{\mathcal{J}}$ then $i \in \mathcal{J}$ by definition. Since also $\tilde{\alpha}_i < w_i \tilde{\lambda}_{\mathcal{J}}$ implies $\mu_i < 0$, hence $i \notin \mathcal{J}$, it follows that we can take $\mathcal{I} = \mathcal{J}$ and $\tilde{\lambda} = \tilde{\lambda}_{\mathcal{J}}$. $\qquad \square$

APPENDIX D.2 DIFFERENTIALS AND THE PROXIMAL OPERATOR OF THE INNER REGULARIZER

We present here expansions of derivations and the proximal operator of the inner regularizer $g^*_{\varepsilon,\delta}$ defined in (5.1). These are needed for the numerical realisation of our method for the experiments of Section 5. Readily, we calculate the partial derivatives of $g^*_{\varepsilon,\delta}$ as (note that each $y_i \in \mathbb{R}^2$)

$$
(\text{D.7}) \qquad \partial_{y_j} g^*_{\varepsilon,\delta}(y) = \partial_{y_j}\left[\max\{0, \frac{1}{3\varepsilon}(\|y_j\| - \alpha_0)^3\} + \frac{\delta}{2}\|y_j\|^2 \right]
$$

$$
= \delta y_j + \begin{cases} 0, & \text{if } \|y_j\| < \alpha_0 \\ \frac{1}{\varepsilon}\frac{y_j}{\|y_j\|}(\|y_j\| - \alpha_0)^2, & \text{otherwise.} \end{cases}
$$

Likewise, the Hessian $\nabla^2_y g^*_{\varepsilon,\delta}(y)$ is given by the entries $\partial_{y_j}\partial_{y_i} g^*_{\varepsilon,\delta}(y) = 0$ for $i \neq j$ and

$$
\partial^2_{y_j} g^*_{\varepsilon,\delta}(y) = \delta\,\mathrm{Id}_2 + \begin{cases} 0, & \text{if } \|y_j\| < \alpha_0 \\ \partial_{y_j}\left[\frac{1}{\varepsilon}\frac{y_j}{\|y_j\|}(\|y_j\| - \alpha_0)^2\right], & \text{otherwise} \end{cases}
$$

$$
= \delta\,\mathrm{Id}_2 + \begin{cases} 0, & \text{if } \|y_j\| < \alpha_0 \\ \frac{1}{\varepsilon}\frac{\mathrm{Id}_2 - y_j y_j^T\|y_j\|^{-2}}{\|y_j\|}(\|y_j\| - \alpha_0)^2 + \frac{1}{\varepsilon}\frac{2y_j y_j^T}{\|y_j\|^2}(\|y_j\| - \alpha_0), & \text{otherwise.} \end{cases}
$$

Lemma D.3. *Let $g^*_{\varepsilon,\delta}$ be defined by (5.1) for some $\varepsilon, \delta > 0$. Moreover, let $v \in \mathbb{R}^{2n_1 n_2}$. Then for any $\tau > 0$ and $j \in \{1, \ldots, n_1 n_2\}$, we have*

$$
[\mathrm{prox}_{\tau g^*_{\varepsilon,\delta}}(v)]_j = \begin{cases} \frac{1}{1+\tau\delta}v_j, & \text{if } \|v_j\| < (1+\tau\delta)\alpha_0 \\ \frac{2\alpha_0 - \varepsilon(\tau^{-1}+\delta) + \sqrt{\varepsilon^2(\tau^{-1}+\delta)^2 + 4\varepsilon(\tau^{-1}\|v_j\| - (\tau^{-1}+\delta)\alpha_0)}}{2\|v_j\|}v_j, & \text{otherwise.} \end{cases}
$$

Proof. Let $y := \mathrm{prox}_{\tau g^*_{\varepsilon,\delta}}(v)$. Let $j \in \{1, \ldots, n_1 n_2\}$. Suppose first that $\|y_j\| < \alpha_0$ Using (D.7) and the the definition of proximal operator, give the characterisation $y_j - v_j + \tau\delta y_j = 0$, i.e., $y_j = \frac{1}{1+\tau\delta}v_j$. Thus $\|v_j\| < (1+\tau\delta)\alpha_0$.

Suppose then that $\|y_j\| \geq \alpha_0$. Again the partial derivative formula (D.7) and the definition of the proximal operator give the characterisation

$$
y_j - v_j + \tau\left(\delta y_j + \frac{1}{\varepsilon}\frac{y_j}{\|y_j\|}(\|y_j\| - \alpha_0)^2\right) = 0.
$$

This is to say $y_j = t v_j$ for some $t > 0$ satisfying

$$
t\left(1 + \tau\left(\delta + \frac{1}{\varepsilon}\frac{1}{t\|v_j\|}(t\|v_j\| - \alpha_0)^2\right)\right) = 1,
$$

which translates into the second order equation in t.

$$
\|v_j\|t^2 + (\varepsilon(\tau^{-1}+\delta) - 2\alpha_0)t + \alpha_0^2\|v_j\|^{-1} - \varepsilon\tau^{-1} = 0.
$$

This is solved by

$$
t = \frac{2\alpha_0 - \varepsilon(\tau^{-1}+\delta) \pm \sqrt{\varepsilon^2(\tau^{-1}+\delta)^2 + 4\varepsilon(\tau^{-1}\|v_j\| - (\tau^{-1}+\delta)\alpha_0)}}{2\|v_j\|}.
$$

Only the positive choice of sign satisfies the ansatz $\|y_j\| \geq \alpha_0$, i.e., $t\|v_j\| \geq \alpha_0$. $\qquad\square$

REFERENCES

[1] S. Angerhausen, *Stochastic Primal-Dual Proximal Splitting Method for Risk-Averse Optimal Control of PDEs*, PhD thesis, University Duisburg–Essen, 2022, doi:10.17185/duepublico/78165.

[2] M. A. Belzunce, High-Resolution Heterogeneous Digital PET [18F]FDG Brain Phantom based on the BigBrain Atlas, 2018, doi:10.5281/zenodo.1190598.

[3] L. Calatroni, C. Cao, J. C. De Los Reyes, C. B. Schönlieb, and T. Valkonen, Bilevel approaches for learning of variational imaging models, in *Variational Methods in Imaging and Geometric Control*, volume 18 of Radon Series on Computational and Applied Mathematics, 2016, 252–290, doi:10.1515/9783110430394-008, arXiv:1505.02120.

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

[5] 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.

[6] T. Chen, Y. Sun, and W. Yin, A single-timescale stochastic bilevel optimization method, 2021, arXiv:2102.04671.

[7] J. Chung, M. Chung, S. Gazzola, and M. Pasha, Efficient learning methods for large-scale optimal inversion design, *Numerical Algebra, Control and Optimization* 14 (2024), 137–159, doi:10.3934/naco.2022036, https://www.aimsciences.org/article/id/63930a2980e8bc15ce45a91f.

[8] F. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, 1990, doi:10.1137/1.9781611971309.

[9] C. Clason and T. Valkonen, *Introduction to Nonsmooth Analysis and Optimization*, 2020. To appear.

[10] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau, A framework for bilevel optimization that enables stochastic and global variance reduction algorithms, 2022, arXiv:2201.13409.

[11] J. C. De Los Reyes and C. B. Schönlieb, Image denoising: Learning noise distribution via PDE-constrained optimization, *Inverse Problems and Imaging* 7 (2013), 1183–1214.

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

[13] J. C. De Los Reyes and D. Villacís, Optimality conditions for bilevel imaging learning problems with total variation regularization, *SIAM Journal on Imaging Sciences* 15 (2022), 1646–1689, doi:10.1137/21m143412x, arXiv:2107.08100.

[14] A. L. Dontchev and R. T. Rockafellar, *Implicit Functions and Solution Mappings: A View from Variational Analysis*, Springer Series in Operations Research and Financial Engineering, Springer, 2014, doi:10.1007/978-1-4939-1037-3.

[15] M. J. Ehrhardt, S. Gazzola, and S. J. Scott, On optimal regularization parameters via bilevel learning, 2023, arXiv:2305.18394.

[16] M. J. Ehrhardt and L. Roberts, Inexact derivative-free optimization for bilevel learning, *Journal of mathematical imaging and vision* 63 (2021), 580–600, doi:10.1007/s10851-021-01020-8.

[17] R. Franzen, Kodak lossless true color image suite, PhotoCD PCD0992. Lossless, true color images released by the Eastman Kodak Company, 1999, http://rok.us/graphics/kodak/.

[18] S. Ghadimi and M. Wang, Approximation methods for bilevel programming, 2018, arXiv:1802.02246.

[19] G. H. Golub and C. F. V. Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 1996.

[20] W. L. Hare and A. S. Lewis, Identifying active constraints via partial smoothness and prox-regularity, *Journal of Convex Analysis* 11 (2004), 251–266.

[21] 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.

[22] B. He and X. Yuan, On the $O(1/n)$ convergence rate of the Douglas–Rachford alternating direction method, *SIAM Journal on Numerical Analysis* 50 (2012), 700–709, doi:10.1137/110836936.

[23] M. Hintermüller, C. N. Rautenberg, T. Wu, and A. Langer, Optimal selection of the regularization function in a weighted total variation model. Part II: Algorithm, its analysis and numerical tests, *Journal of Mathematical Imaging and Vision* (2017), doi:10.1007/s10851-017-0736-2.

[24] M. Hintermüller and T. Wu, Bilevel optimization for calibrating point spread functions in blind deconvolution, *Inverse Problems and Imaging* 9 (2015), 1139–1169, doi:10.3934/ipi.2015.9.1139.

[25] G. Holler, K. Kunisch, and R. C. Barnard, A bilevel approach for parameter learning in inverse problems, *Inverse Problems* 34 (2018), 115012, doi:10.1088/1361-6420/aade77.

[26] M. Hong, H. T. Wai, Z. Wang, and Z. Yang, A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic, 2020, arXiv:2007.05170.

[27] A. D. Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, Springer Monographs in Mathematics, Springer, 2017, doi:10.1007/978-3-319-64277-2.

[28] B. Jensen and T. Valkonen, A nonsmooth primal-dual method with interwoven PDE constraint solver, *Computational Optimization and Applications* 89 (2024), 115–149, doi:10.1007/s10589-024-00587-3, arXiv:2211.04807.

[29] K. Ji and Y. Liang, Lower bounds and accelerated algorithms for bilevel optimization, *Journal of Machine Learning Research* 23 (2022), 1–56.

[30] K. Ji, J. Yang, and Y. Liang, Bilevel optimization: Convergence analysis and enhanced design, in *International Conference on Machine Learning*, 2021, 4882–4892.

[31] K. Kunisch and T. Pock, A bilevel optimization approach for parameter learning in variational models, *SIAM Journal on Imaging Sciences* 6 (2013), 938–983, doi:10.1137/120882706.

[32] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak, A fully first-order method for stochastic bilevel optimization, in *International Conference on Machine Learning*, 2023, 18083–18113.

[33] J. Li, B. Gu, and H. Huang, A fully single loop algorithm for bilevel optimization without hessian inverse, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, 7426–7434.

[34] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu, Bome! bilevel optimization made easy: A simple first-order approach, *Advances in neural information processing systems* 35 (2022), 17248–17262.

[35] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond, 2021, arXiv:2101.11517.

[36] Q. Rebjock and N. Boumal, Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions, *Mathematical Programming* (2024), doi:10.1007/s10107-024-02136-6.

[37] M. S. Salehi, S. Mukherjee, L. Roberts, , and M. J. Ehrhardt, An adaptively inexact first-order method for bilevel optimization with application to hyperparameter learning, 2024, arXiv:2308.10098.

[38] F. Sherry, M. Benning, J. C. De Los Reyes, M. J. Graves, G. Maierhofer, G. Williams, C. B. Schönlieb, and M. J. Ehrhardt, Learning the sampling pattern for MRI, *IEEE Transactions on Medical Imaging* 39 (2020), 4310–4321, doi:10.1109/tmi.2020.3017353.

[39] P. Sonneveld, CGS, A fast Lanczos-type solver for nonsymmetric linear systems, *SIAM Journal on Scientific and Statistical Computing* 10 (1989), 36–52, doi:10.1137/0910004.

[40] E. Suonperä and T. Valkonen, Linearly convergent bilevel optimization with single-step inner methods, *Computational Optimization and Applications* (2023), doi:10.1007/s10589-023-00527-7, arXiv:2205.04862.

[41] E. Suonperä and T. Valkonen, Codes for "Single-loop methods for bilevel parameter learning in inverse imaging", 2024, doi:10.5281/zenodo.15006162.

[42] 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.

[43] T. Valkonen, Preconditioned proximal point methods and notions of partial subregularity, *Journal of Convex Analysis* 28 (2021), 251–278, arXiv:1711.05123.

[44] T. Valkonen, Regularisation, optimisation, subregularity, *Inverse Problems* 37 (2021), 045010, doi:10.1088/1361-6420/abe4aa, arXiv:2011.07575.

[45] H. A. Van der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, *SIAM Journal on Scientific Computation* 13 (1992), 631–644.

[46] J. Yang, K. Ji, and Y. Liang, Provably faster algorithms for bilevel optimization, *Advances in Neural Information Processing Systems* 34 (2021), 13670–13682.