LINEARLY CONVERGENT BILEVEL OPTIMIZATION WITH SINGLE-STEP INNER METHODS

Ensio Suonperä* Tuomo Valkonen†

Abstract We propose a new approach to solving bilevel optimization problems, intermediate between solving full-system optimality conditions with a Newton-type approach, and treating the inner problem as an implicit function. The overall idea is to solve the full-system optimality conditions, but to precondition them to alternate between taking steps of simple conventional methods for the inner problem, the adjoint equation, and the outer problem. While the inner objective has to be smooth, the outer objective may be nonsmooth subject to a prox-contractivity condition. We prove *linear convergence* of the approach for combinations of gradient descent and forward-backward splitting with exact and inexact solution of the adjoint equation. We demonstrate good performance on learning the regularization parameter for anisotropic total variation image denoising, and the convolution kernel for image deconvolution.

1 INTRODUCTION

Two general approaches are typical for the solution of the bilevel optimization problem

$$\min_{\alpha\in\mathscr{A}} J(S_u(\alpha)) + R(\alpha) \quad \text{with} \quad S_u(\alpha) \in \operatorname*{arg\,min}_{u\in U} F(u;\alpha) \tag{1.1}$$

in Hilbert spaces $\mathscr{A}$ and U. The first, familiar from the treatment of general mathematical programming with equilibrium constraints (MPECs), is to write out the Karush–Kuhn–Tucker conditions for the whole problem in a suitable form, and to apply a Newton-type method or other nonlinear equation solver to them [4, 1, 25, 36, 18].

The second approach, common in the application of (1.1) to inverse problems and imaging [24, 14, 37, 32, 30, 45, 6], treats the solution mapping S_u as an implicit function. Thus it is necessary to (i) on each outer iteration k solve the inner problem $\min_u F(u;\alpha^k)$ near-exactly using an optimization method of choice; (ii) solve an adjoint equation to calculate the gradient of the solution mapping; and (iii) use another optimization method of choice on the outer problem $\min_\alpha J(S_u(\alpha))$ with the knowledge of $S_u(\alpha^k)$ and $\nabla S_u(\alpha^k)$. The inner problem is therefore generally assumed to have a unique solution, and the solution map to be differentiable. An algorithm for nonsmooth inner problems has been developed in [39], while [17] rely on proving directional Bouligand differentiability for otherwise nonsmooth problems.

The challenge of the first "whole-problem" approach is to scale it to large problems, typically involving the inversion of large matrices. The difficulty with the second "implicit function" approach is that the inner problem needs to be solved several times, which can be expensive. Solving the adjoint

This research has been supported by the Academy of Finland grants 314701, 320022, and 345486.

*Department of Mathematics and Statistics, University of Helsinki, Finland. ensio.suonpera@helsinki.fi

†ModeMat, Escuela Politécnica Nacional, Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland. tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

equation also requires matrix inversion. The variant in [23] avoids this through derivative-free methods for the outer problem. It also solves the inner problem to a low but controlled accuracy.

In this paper, by preconditioning the implicit-form first-order optimality conditions, we develop an intermediate approach more efficient than the aforementioned, as we demonstrate in the numerical experiments of Section 4. It can be summarized as (i) take only *one* step of an optimization method on the inner problem, (ii) perform a *cheap* operation to advance towards the solution of the adjoint equation, and, finally, (iii) using this approximate information, take one step of an optimization method for the outer problem. Repeat.

The preconditioning, which we introduce in detail in Section 2, is based on insight from the derivation of the primal-dual proximal splitting of [7] as a preconditioned proximal point method [29, 49, 12]. We write the optimality conditions for (1.1) as the inclusion $0 \in H(x)$ for a set-valued H, where $x = (u, p, \alpha)$ for an adjoint variable p. The basic proximal point method then iteratively solves x^{k+1} from

$$0 \in H(x^{k+1}) + (x^{k+1} - x^k).$$

This can be as expensive as solving the original optimality condition. The idea then is to introduce a preconditioning operator M that *decouples* the components of x—in our case u, p and α—such that each component can be solved in succession from

$$0 \in H(x^{k+1}) + M(x^{k+1} - x^k).$$

Gradient steps can be handled through nonlinear preconditioning [49, 12], as we will see in Section 2 when we develop the approach in detail along with two more specific algorithms, the FEFB (Forward–Exact–Forward-Backward) and the FIFB (Forward–Inexact–Forward-Backward). In Section 3 we prove their local linear convergence under a second-order growth condition on the composed objective $J \circ S_u$, and other more technical conditions. The proof is based on the "testing" approach developed in [49] and also employed extensively in [12, 50]. Finally, we evaluate the numerical performance of the proposed schemes on imaging applications in Section 4, specifically the learning of a regularization parameter for total variation denoising, and the convolution kernel for deblurring. Since the purpose of these experiments is a simple performance comparison between different algorithms, instead of real applications, we only use a single training sample of various dimensions, as explained in Section 4.

Intermediate approaches, some reminiscent of ours, have recently also been developed in the machine learning community. Our approach, however, allows a non-smooth function R in the outer problem (1.1). Moreover, to our knowledge, our work is the first to show linear convergence for a fully "single-loop" algorithm. To be more precise, the STABLE [9], TTSA [33], FLSA [38], MRBO, VRBO [53], and SABA [13] are "single-loop" algorithms such as ours, taking only a single step towards the solution of the inner problem on each outer iteration. The STABLE requires solving the adjoint equation exactly, as does our first approach, the FEFB. The others use a Neumann series approximation for the adjoint equation. Our second approach, the FIFB, takes a simple step reminiscent of gradient descent for the adjoint equation. The TSSA and STABLE obtain sublinear convergence of the outer iterates $\{\alpha^k\}_{k\in\mathbb{N}}$ assuming strong convexity (second-order growth) of both the inner and outer objective. For the SABA similar linear convergence is claimed with the outer strong convexity replaced by a Polyak-Łojasiewicz inequality. Without either of those assumptions, the theoretical results on the aforementioned methods from the literature are much weaker, and generally only show various forms of "stall" of the iterates at a sublinear rate, or the ergodic convergence of the gradient $\nabla_\alpha[J \circ S_u](\alpha^k)$ of the composed objective to zero. Such modes of convergence say very little about the convergence of function values to optimum or the iterates to a solution.

In the context of not fully single-loop algorithms, the AID, ITD [35], AccBio [34], and ABA [27] take a fixed (small) number of inner iterations for each outer iteration. The AID and ITD only sublinear convergence of the composed gradient is claimed. For the ABA and AccBio linear convergence of outer function values is claimed under strong convexity of both the inner and outer objectives.

FUNDAMENTALS AND APPLICATIONS

Fundamentals of MPECs and bilevel optimization are treated in the books [40, 2, 19, 21]. An extensive literature review up to 2018 can be found in [20], and recent developments in [46]. Optimality conditions for bilevel problems, both necessary and sufficient, are developed in, e.g., [54, 55, 22, 41, 3]. A more limited type of "bilevel" problems only constrains α to lie in the set of minimisers of another problem. Algorithms for such problems are treated in [43, 44].

Bilevel optimization has been used for *learning* regularization parameters and forward operators for inverse imaging problems. With total variation regularization in the inner problem, the parameter learning problem in its most basic form reads [14]

$$\min_{\alpha} \ \frac{1}{2}\|S_u(\alpha) - b\|^2 + R(\alpha) \quad \text{with} \quad S_u(\alpha) = \operatorname*{arg\,min}_{u \in U} \frac{1}{2}\|A_\alpha u - z\|^2 + \alpha_1 \|\nabla u\|_1.$$

This problem finds the best possible α for reconstructing the "ground truth" image b from the measurement data z, which may be noisy and possibly transformed and only partially known through the *forward operator* A_α, mapping images to measurements. To generalize to multiple images, the outer problem would sum over them and corresponding inner problems [6]. Multi-parameter regularization is discussed in [16], and natural conditions for $\alpha > 0$ in [15].[1] In other works, the forward operator A_α is learned for blind image deblurring [31] or undersampling in magnetic resonance imaging [45]. In [37] regularization kernels are learned, while [17, 8] study the learning of optimal discretisation schemes. To circumvent the non-differentiability of S_u, [42] replace the inner problem with a fixed number of iterations of an algorithm. Their approach has connections to the learning of deep neural networks.

Bilevel problems can also be seen as leader–follower or *Stackelberg games*: the outer problem or agent leads by choosing α, and the inner agent reacts with the best possible u for that α. Multiple-agent *Nash equilibria* may also be modeled as bilevel problems. Both types of games can be applied to financial markets and resource use planning; we refer to the the aforementioned books [40, 2, 19, 21] for specific examples.

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^*.

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 $\nabla E(x) := G'(x)^* \in \mathbb{L}(Y;X)$. 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., $\nabla_u F(u,\alpha)$. For convex $R : X \to \overline{\mathbb{R}}$, we write $\operatorname{dom} R$ 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 operator as $\operatorname{prox}_R(x) := \arg\min_z \frac{1}{2}\|z - x\|^2 + R(z) = (\mathrm{Id} + \partial R)^{-1}(x)$.

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 $M \in \mathbb{L}(X;X)$ we write $\|x\|_M := \sqrt{\langle x, x\rangle_M} := \sqrt{\langle Mx, x\rangle}$. Pythagoras' or *three-point identity* then states

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

[1] An error in [15, Lemma 10] requires some conditions therein to be taken "in the limit" as $t \searrow 0$.

We extensively use Young's inequality

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

We sometimes apply operations on $x \in X$ to all elements of a set $A \subset X$, writing $\langle x+A, z\rangle := \{\langle x+a, z\rangle \mid a \in A\}$, and for $B \subset \mathbb{R}$, writing $B \ge c$ if $b \ge c$ for all $b \in B$.

2 PROPOSED METHODS

We now present our proposed methods for (1.1). They are based on taking a single gradient descent step for the inner problem, and using forward-backward splitting for the outer problem. The two methods differ on how an "adjoint equation" is handled. We present the algorithms and assumptions required to prove their convergence in Sections 2.2 and 2.3 after deriving optimality conditions and the adjoint equation in Section 2.1. We prove convergence in Section 3.

2.1 OPTIMALITY CONDITIONS

Suppose $u \mapsto F(u;\alpha) \in C^2(U)$ is proper, coercive, and weakly lower semicontinuous for each *outer variable* $\alpha \in \operatorname{dom} R \subset \mathscr{A}$. Then the direct method of the calculus of variations guarantees the inner problem $\min_u F(u;\alpha)$ to have a solution. If, further, $u \mapsto F(u;\alpha)$ is strictly convex, the solution is unique so that the solution mapping S_u from (1.1) is uniquely determined.

Suppose further that $F, \nabla F$ and S_u are Fréchet differentiable. Writing $T(\alpha) := (S_u(\alpha), \alpha)$, Fermat's principle and $S_u(\tilde\alpha) \in \arg\min_u F(u;\tilde\alpha)$ then show that

$$[\nabla_u F \circ T](\alpha) = \nabla_u F(S_u(\alpha);\alpha) = 0 \tag{2.1}$$

for α near $\tilde\alpha$. Therefore, the chain rule for Fréchet differentiable functions yields

$$0 = \nabla_\alpha[\nabla_u F \circ T](\alpha) = \nabla_\alpha S_u(\alpha)\nabla_u^2 F(T(\alpha)) + \nabla_{\alpha u} F(T(\alpha)).$$

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

$$0 = p\nabla_u^2 F(u,\alpha) + \nabla_{\alpha u} F(u,\alpha). \tag{2.2}$$

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

$$S_p(u,\alpha) := -\nabla_{\alpha u} F(u;\alpha)\left(\nabla_u^2 F(u;\alpha)\right)^{-1}. \tag{2.3}$$

We will later make assumptions that ensure that S_p is well-defined.

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 $p \in \mathbb{L}(U;\mathscr{A})$, but we will need p to lie in an inner product space. Assuming $\mathscr{A}$ to be a *separable* Hilbert space, we introduce such structure

$$P = (\mathbb{L}(U;\mathscr{A}), \langle\!\langle \cdot, \cdot \rangle\!\rangle) \tag{2.4a}$$

by using a countable orthonormal basis $\{\varphi_i\}_{i\in I}$ of $\mathscr{A}$ to define the inner product

$$\langle\!\langle p_1, p_2\rangle\!\rangle := \sum_{i\in I}\langle p_1^*\varphi_i, p_2^*\varphi_i\rangle = \sum_{i\in I}\langle \varphi_i, p_1 p_2^*\varphi_i\rangle. \quad (p_1, p_2 \in \mathbb{L}(U;\mathscr{A})). \tag{2.4b}$$

We briefly study this inner product and the induced norm $|||\cdot|||$ in Appendix B.

Algorithm 2.1 Forward–exact–forward-backward (FEFB) method

Require: Functions R, J, F as in Assumption 2.2. Step length parameters $\tau, \sigma > 0$.

1: Pick an initial iterate $(u^0, \alpha^0) \in U \times \mathscr{A}$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad u^{k+1} := u^k - \tau \nabla_u F(u^k; \alpha^k)$ ▷ inner gradient step
4: $\quad p^{k+1} := -\nabla_{\alpha u} F(u^{k+1}; \alpha^k) \left(\nabla_u^2 F(u^{k+1}; \alpha^k)\right)^{-1}$ ▷ adjoint solution
5: $\quad \alpha^{k+1} := \operatorname{prox}_{\sigma R}\left(\alpha^k - \sigma(p^{k+1} \nabla_u J(u^{k+1}))\right)$ ▷ outer forward-backward step
6: **end for**

By the sum rule for Clarke subdifferentials (denoted ∂_C) and their compatibility with convex subdifferentials and Fréchet differentiable functions [10], 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 subdifferentials then furnishes the necessary optimality condition

$$0 \in \nabla_\alpha (J \circ S_u)(\widehat{\alpha}) + \partial R(\widehat{\alpha}) = \nabla_\alpha S_u(\widehat{\alpha}) \nabla_u J(S_u(\widehat{\alpha})) + \partial R(\widehat{\alpha}). \tag{2.5}$$

We combine (2.1), (2.2) and (2.5) as the inclusion

$$0 \in H(\widehat{u}, \widehat{p}, \widehat{\alpha}) \tag{2.6}$$

with

$$H(u, p, \alpha) := \begin{pmatrix} \nabla_u F(u; \alpha) \\ p \nabla_u^2 F(u; \alpha) + \nabla_{\alpha u} F(u; \alpha) \\ p \nabla_u J(u) + \partial R(\alpha) \end{pmatrix} \quad \text{for all} \quad u \in U, p \in P, \alpha \in \mathscr{A}. \tag{2.7}$$

This is the optimality condition that our proposed methods, presented in Sections 2.2 and 2.3, attempt to satisfy. We generally abbreviate

$$x = (u, p, \alpha), \quad \widehat{x} = (\widehat{u}, \widehat{p}, \widehat{\alpha}), \quad \text{etc.}$$

2.2 ALGORITHM: FORWARD–EXACT–FORWARD-BACKWARD

Our first strategy for solving (2.6) takes just a *single* gradient descent step for the inner problem, solves the adjoint equation exactly, and then takes a forward-backward step for the outer problem. We call this Algorithm 2.1 the FEFB (forward–exact–forward-backward).

Using H defined in (2.7), Algorithm 2.1 can be written implicitly as solving

$$0 \in H_{k+1}(x^{k+1}) + M(x^{k+1} - x^k) \tag{2.8}$$

for $x^{k+1} = (u^{k+1}, p^{k+1}, \alpha^{k+1})$, where, with $x = (u, p, \alpha)$,

$$H_{k+1}(x) := H(x) + \begin{pmatrix} \nabla_u F(u^k; \alpha^k) - \nabla_u F(u; \alpha) \\ p \nabla_u^2 F(u; \alpha^k) + \nabla_{\alpha u} F(u; \alpha^k) - \left(p \nabla_u^2 F(u; \alpha) + \nabla_{\alpha u} F(u; \alpha)\right) \\ 0 \end{pmatrix}, \tag{2.9a}$$

and the *preconditioning operator* $M \in \mathbb{L}(U \times P \times \mathscr{A}; U \times P \times \mathscr{A})$ is

$$M := \operatorname{diag}(\tau^{-1} \operatorname{Id}, 0, \sigma^{-1} \operatorname{Id}). \tag{2.9b}$$

The "nonlinear preconditioning" applied to H to construct H_{k+1} shifts iterate indices such that a forward step is performed instead of a proximal step; compare [49, 12].

We next state essential structural, initialisations, 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 $A \subset \operatorname{dom} R$) if there exist $C_R > 0$ and a neighborhood $A \subset \operatorname{dom} R$ of $\widehat{\alpha}$ such that, for all $\alpha \in A$,

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

We verify Assumption 2.1 for some common cases in Appendix A. When applying the assumption to to $\widehat{\alpha}$ satisfying (2.6), we will take $q = -\widehat{p}\nabla_u J(\widehat{u}) \in \partial R(\widehat{\alpha})$. Then $D_{\sigma R}(\widehat{\alpha}) = 0$ by standard properties of proximal mappings. The results for nonsmooth functions in Appendix A in that case forbid *strict complementarity*. In particular, for $R = \beta\|\cdot\|_1 + \delta_{[0,\infty)^n}$ we need to have $q \in (\beta, \dots, \beta)$, and for $R = \delta_C$ for a convex set C, we need to have $q = 0$. Intuitively, this restriction serves to forbid the *finite identification* property [28] of proximal-type methods, as $\{\alpha^n\}$ cannot converge too fast in our 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 FEFB. It collects conditions related to step lengths, initialization, and the problem functions F, J, and R. For a constant $c > 0$ to be determined by the assumption, we introduce the *testing operator*

$$Z := \operatorname{diag}(\varphi_u \operatorname{Id}, \operatorname{Id}, \operatorname{Id}). \tag{2.10}$$

The idea, introduced in [49] and further explained in [12], is to *test* the algorithm-defining inclusion (2.8) with the linear functional $\langle Z\,\cdot, x^{k+1} - \widehat{x}\rangle$ to obtain a descent estimate with respect to the ZM-norm. The operator Z encodes component-specific scalings and convergence rates, although we do not exploit the latter in this manuscript.

Assumption 2.2. We assume that U is a Hilbert space, $\mathscr{A}$ a separable Hilbert space, and treat the adjoint variable $p \in \mathbb{L}(U; \mathscr{A})$ as an element of the inner product space P defined in (2.4). Let $R : \mathscr{A} \to \overline{\mathbb{R}}$ and $J : U \to \mathbb{R}$ be convex, proper, and lower semicontinuous, and assume the same from $F(\,\cdot\,, \alpha) \in C^2(U)$ for all $\alpha \in \operatorname{dom} R$. Pick $(\widehat{u}, \widehat{p}, \widehat{\alpha}) \in H^{-1}(0)$ and let $\{(u^m, p^m, \alpha^m)\}_{m\in\mathbb{N}}$ be generated by Algorithm 2.1 for a given initial iterate $(u^0, p^0, \alpha^0) \in U \times P \times \operatorname{dom} R$. For a given $r, r_u > 0$ we suppose that

(i) The relative initialization bound $\|u^1 - S_u(\alpha^0)\| \leq C_u\|\alpha^0 - \widehat{\alpha}\|$ holds for some $C_u > 0$.

(ii) There exists in $B(\widehat{\alpha}, 2r) \cap \operatorname{dom} R$ a continuously Fréchet-differentiable and L_{S_u}-Lipschitz inner problem solution mapping $S_u : \alpha \mapsto S_u(\alpha) \in \arg\min F(\,\cdot\,; \alpha)$.

(iii) $F(\widehat{u};\,\cdot\,)$ is Lipschitz continuously differentiable with factor $L_{\nabla F,\widehat{u}} > 0$, and $\gamma_F \cdot \operatorname{Id} \leq \nabla_u^2 F(u;\alpha) \leq L_F \cdot \operatorname{Id}$ for all $(u, \alpha) \in B(\widehat{u}, r_u) \times (B(\widehat{\alpha}, r) \cap \operatorname{dom} R)$ for some $\gamma_F, L_F > 0$. Moreover, $(u, \alpha) \mapsto \nabla_u^2 F(u;\alpha)$ and $(u, \alpha) \mapsto \nabla_{\alpha u} F(u;\alpha) \in P$ are Lipschitz in $B(\widehat{u}, r_u) \times (B(\widehat{\alpha}, r) \cap \operatorname{dom} R)$ with factors $L_{\nabla^2 F}$ and $L_{\nabla_{\alpha u} F}$, where we equip $U \times \mathscr{A}$ with the norm $(u, \alpha) \mapsto \|u\|_U + \|\alpha\|_{\mathscr{A}}$.

(iv) The inner step length $\tau \in (0, 2\kappa/L_F]$ for some $\kappa \in (0, 1)$.

(v) The outer fitness function J is Lipschitz continuously differentiable with factor $L_{\nabla J}$, and $\gamma_\alpha \cdot \operatorname{Id} \leq \nabla_\alpha^2(J \circ S_u) \leq L_\alpha \cdot \operatorname{Id}$ 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 $B(\widehat{\alpha}, r) \cap \operatorname{dom} R$ for some $C_R \geq 0$.

(vi) The constants $\varphi_u, C_u > 0$ satisfy

$$\gamma_F(L_{\nabla J} N_p + L_{S_p} N_{\nabla J}) C_u + \frac{L_{\nabla F,\widehat{u}}^2}{(1-\kappa)} \varphi_u < \gamma_F \gamma_\alpha,$$

where

$$N_{\nabla_{\alpha u}F} := \max_{\substack{u\in B(\widehat{u},r_u),\\ \alpha\in B(\widehat{\alpha},2r)\cap\operatorname{dom} R}} |||\nabla_{\alpha u}F(u,\alpha)|||, \qquad L_{S_p} := \gamma_F^{-2}L_{\nabla^2 F}N_{\nabla_{\alpha u}F} + \gamma_F^{-1}L_{\nabla_{\alpha u}F},$$

$$N_{\nabla J} := \max_{\alpha\in B(\widehat{\alpha},r)\cap\operatorname{dom} R} \|\nabla_u J(S_u(\alpha))\|, \qquad N_{\nabla S_u} := \max_{\alpha\in B(\widehat{\alpha},r)\cap\operatorname{dom} R} |||\nabla_\alpha S_u(\alpha)|||, \text{ and}$$

$$N_p := N_{\nabla S_u} + Cr \text{ with } C = L_{S_p}C_u.$$

(vii) The outer step length σ fulfills

$$0 < \sigma \le \frac{(C_F-1)C_u}{(L_{S_u}+C_FC_u)C_\alpha} \quad \text{where} \quad \begin{cases} C_F := \sqrt{1+2\tau\gamma_F(1-\kappa)}, & \text{and} \\ C_\alpha := (N_pL_{\nabla J} + N_{\nabla J}L_{S_p})C_u + L_\alpha + C_R. \end{cases}$$

(viii) The initial iterates u^0 and α^0 are such that the distance-to-solution

$$r_0 := \sqrt{\sigma\varphi_u\tau^{-1}\|u^0-\widehat{u}\|^2 + \|\alpha^0-\widehat{\alpha}\|^2} = \sqrt{\sigma}\|x^0-\widehat{x}\|_{ZM}$$

satisfies

$$r_0 \le r \quad \text{and} \quad r_0\max\{2L_{S_u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}(1+\tau L_F) + \tau L_{\nabla F,\widehat{u}}\} \le r_u.$$

Remark 2.3 (Interpretation). Part (i) of Assumption 2.2 ensures that the initial inner problem iterate is good *relative* to the outer problem iterate. If u^1 solves the inner problem for α^0, (i) holds for any $C_u > 0$. Therefore, (i) 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.

Part (ii) ensures that the inner problem solution map exists and is well-behaved; we discuss it more in the next Remark 2.4.

Parts (iii) and (v) are second order growth and boundedness conditions, standard in smooth optimization. The nonsmooth R is handled through the prox-σ-contractivity assumption. If S_u is twice Fréchet differentiable, the product and the chain rules establish

$$\langle h, \nabla_\alpha^2(J\circ S_u)(\alpha)h\rangle = \langle S_u'(\alpha)h, \nabla_u^2 J(S_u(\alpha))S'(\alpha)h\rangle + \langle \nabla_u J(S_u(\alpha)), S_u''(\alpha)(h,h)\rangle \quad (h\in\mathscr{A}).$$

If $R = 0$, first-order optimality conditions establish $\nabla_u J(S_u(\widehat{\alpha})) = 0$. Therefore, if, further, J is strongly convex and $S_u'(\widehat{\alpha})$ is invertible, $\gamma\cdot\mathrm{Id} \le \nabla_\alpha^2(J\circ S_u)(\widehat{\alpha})$ for some $\gamma > 0$. Then additional continuity assumptions establish the positivity required in (v) in a neighbourhood of $\widehat{\alpha}$. It is also possible to further develop the condition to not depend on the solution mapping at all.

Dependent on R, (v) may restrict the outer step length parameter σ. Part (iii) ensures that $u \mapsto \nabla_u^2 F(u;\alpha)$ is invertible and S_p is well-defined. We will see in Lemma 3.3 that the radius r_u is sufficiently large that $\alpha\in B(\widehat{\alpha},r)$ implies $S_u(\alpha)\in B(\widehat{u},r_u)$. Part (v) implies that $\alpha\mapsto\nabla_\alpha(J\circ S_u)(\alpha)$ is Lipschitz in $B(\widehat{\alpha},r)$.

Part (iv) is a standard step length condition for the inner problem while (vii) is a step length condition for the outer problem. It depends on several constants defined in the more technical part (vi). *We can always satisfy the inequality in* (vi) *by good relative initialisation (small $C_u > 0$), as discussed above, and taking the testing parameter φ_u small.* According to the local initialization condition (viii), the latter can be done if the initialial iterates are close to a solution $(\widehat{u},\widehat{\alpha})$ of the entire problem, or if $r_u > 0$ can be be taken arbitrarily large. If we can take both $r > 0$ and $r_u > 0$ arbitrarily large, we obtain *global convergence.*

Remark 2.4 (Existence and differentiability of the solution map). Suppose F is twice continuously differentiable in both variables, and that $\gamma_F\cdot\mathrm{Id} \le \nabla_u^2 F(u;\alpha)$ for all $u\in B(\widehat{u},r_u)$ and $\alpha\in B(\widehat{\alpha},2r)\cap\operatorname{dom} R$ for some $\gamma_F > 0$. Then the implicit function theorem shows the existence of a unique continuously differentiable S_u in a neighborhood of any $\alpha\in B(\widehat{\alpha},r)\cap\operatorname{dom} R$. Such an S_u is also Lipschitz in a neighborhood of α; see, e.g., [12, Lemma 2.11]. If $\mathscr{A}$ is finite-dimensional, a compactness argument gluing together the neighborhoods then proves Assumption 2.2 (ii).

Algorithm 2.2 Forward–inexact–forward-backward (FIFB) method

Require: Functions R, J, F as in Assumption 2.5. Step length parameters $\tau, \sigma, \theta > 0$.

1: Pick an initial iterate $(u^0, p^0, \alpha^0) \in U \times P \times \mathscr{A}$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad u^{k+1} := u^k - \tau \nabla_u F(u^k; \alpha^k)$ ▷ inner gradient step
4: $\quad p^{k+1} := p^k - \theta \left(p^k \nabla_u^2 F(u^{k+1}; \alpha^k) + \nabla_{\alpha u} F(u^{k+1}; \alpha^k) \right)$ ▷ adjoint step
5: $\quad \alpha^{k+1} := \operatorname{prox}_{\sigma R} \left(\alpha^k - \sigma (p^{k+1} \nabla_u J(u^{k+1})) \right)$ ▷ outer forward-backward step
6: **end for**

2.3 ALGORITHM: FORWARD–INEXACT–FORWARD-BACKWARD

Our second strategy for solving (2.6) modifies the first approach to solve the adjoint variable inexactly, so that no costly matrix inversions are required. Instead we perform an update reminiscent of a gradient step. This approach, which we call the FIFB (forward–inexact–forward-backward) reads as Algorithm 2.2 and has the implicit form

$$\begin{cases} 0 = \tau \nabla_u F(u^k; \alpha^k) + u^{k+1} - u^k \\ 0 = \theta \left(p^k \nabla_u^2 F(u^{k+1}; \alpha^k) + \nabla_{\alpha u} F(u^{k+1}; \alpha^k) \right) + p^{k+1} - p^k \\ 0 \in \sigma (\partial R(\alpha^{k+1}) + p^{k+1} \nabla_u J(u^{k+1})) + \alpha^{k+1} - \alpha^k. \end{cases} \tag{2.11}$$

The implicit form can also be written as (2.8) with

$$H_{k+1}(x) := H(x) + \begin{pmatrix} \nabla_u F(u^k; \alpha^k) - \nabla_u F(u; \alpha) \\ p^k \nabla_u^2 F(u; \alpha^k) + \nabla_{\alpha u} F(u; \alpha^k) - \left(p \nabla_u^2 F(u; \alpha) + \nabla_{\alpha u} F(u; \alpha) \right) \\ 0 \end{pmatrix}, \tag{2.12a}$$

and the preconditioning operator $M \in \mathbb{L}(U \times P \times \mathscr{A}; U \times P \times \mathscr{A})$,

$$M := \operatorname{diag}(\tau^{-1} \operatorname{Id}, \theta^{-1} \operatorname{Id}, \sigma^{-1} \operatorname{Id}). \tag{2.12b}$$

For the testing operator Z we use the structure

$$Z := \operatorname{diag}(\varphi_u \operatorname{Id}, \varphi_p \operatorname{Id}, \operatorname{Id}). \tag{2.13}$$

with the constants $\varphi_u, \varphi_p > 0$ determined in the following assumption. It is the FIFB counterpart of Assumption 2.2 for the FEFB, collecting essential structural, step length, and initialization assumptions.

Assumption 2.5. We assume that U is a Hilbert space, $\mathscr{A}$ a separable Hilbert space, and treat the adjoint variable $p \in \mathbb{L}(U; \mathscr{A})$ as an element of the inner product space P defined in (2.4). Let $R : \mathscr{A} \to \overline{\mathbb{R}}$ and $J : U \to \mathbb{R}$ be convex, proper, and lower semicontinuous, and assume the same from $F(\bullet, \alpha)$ for all $\alpha \in \operatorname{dom} R$. Pick $(\widehat{u}, \widehat{p}, \widehat{\alpha}) \in H^{-1}(0)$ and let $\{(u^m, p^m, \alpha^m)\}_{m \in \mathbb{N}}$ be generated by Algorithm 2.2 for a given initial iterate $(u^0, p^0, \alpha^0) \in U \times P \times \operatorname{dom} R$. For given $r, r_u > 0$ we suppose that

(i) The relative initialization bounds $\|u^1 - S_u(\alpha^0)\| \leq C_u \|\alpha^0 - \widehat{\alpha}\|$ and $\|\!|p^1 - \nabla_\alpha S_u(\alpha^0)|\!\| \leq C_p \|\alpha^0 - \widehat{\alpha}\|$ hold with some constants $C_u > 0$ and $C_p > 0$.

(ii) There exists in $B(\widehat{\alpha}, 2r) \cap \operatorname{dom} R$ a continuously Fréchet-differentiable and L_{S_u}-Lipschitz inner problem solution mapping $S_u : \alpha \mapsto S_u(\alpha) \in \arg\min F(\bullet; \alpha)$.

(iii) $F(\widehat{u}; \bullet)$ is Lipschitz continuously differentiable with factor $L_{\nabla F, \widehat{u}} > 0$, and $\gamma_F \cdot \operatorname{Id} \leq \nabla_u^2 F(u; \alpha) \leq L_F \cdot \operatorname{Id}$ for $u \in B(\widehat{u}, r_u)$ and $\alpha \in B(\widehat{\alpha}, 2r) \cap \operatorname{dom} R$. Moreover, $(u, \alpha) \mapsto \nabla_u^2 F(u; \alpha)$ and $(u, \alpha) \mapsto \nabla_{\alpha u} F(u; \alpha) \in P$ are Lipschitz in $B(\widehat{u}, r_u) \times (B(\widehat{\alpha}, r) \cap \operatorname{dom} R)$ with factors $L_{\nabla^2 F}$ and $L_{\nabla_{\alpha u} F}$, where we equip $U \times \mathscr{A}$ with the norm $(u, \alpha) \mapsto \|u\|_U + \|\alpha\|_{\mathscr{A}}$.

(iv) The inner step length $\tau \in (0, 2\kappa/L_F]$ for some $\kappa \in (0,1)$ whereas the adjoint step length $\theta \in (0, 1/L_F)$.

(v) The outer fitness function J is Lipschitz continuously differentiable with factor $L_{\nabla J}$, and $\gamma_\alpha \cdot \mathrm{Id} \leq \nabla^2_\alpha (J \circ S_u) \leq L_\alpha \cdot \mathrm{Id}$ 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 $B(\widehat{\alpha}, r) \cap \operatorname{dom} R$ for some $C_R \geq 0$.

(vi) The constants $\varphi_u, \varphi_p, C_u > 0$ satisfy

$$\varphi_p \leq \varphi_u \frac{\gamma_F^2(1-\kappa)}{2L_F L_{S_p}}$$

and

$$L_F L_{S_p}\varphi_p + \sqrt{(L_F L_{S_p}\varphi_p)^2 + \gamma_F^2(L_{\nabla J}N_p + L_{S_p}N_{\nabla J})^2 C_u^2} + \frac{L^2_{\nabla F,\widehat{u}}}{(1-\kappa)}\varphi_u < \gamma_F \gamma_\alpha,$$

where

$$N_{\nabla_{\alpha u}F} := \max_{\substack{u \in B(\widehat{u}, r_u),\\ \alpha \in B(\widehat{\alpha}, 2r) \cap \operatorname{dom} R}} |||\nabla_{\alpha u} F(u,\alpha)|||, \qquad L_{S_p} := \gamma_F^{-2} L_{\nabla^2 F} N_{\nabla_{\alpha u}F} + \gamma_F^{-1} L_{\nabla_{\alpha u}F},$$

$$N_{\nabla J} := \max_{\alpha \in B(\widehat{\alpha}, r) \cap \operatorname{dom} R} \|\nabla_u J(S_u(\alpha))\|, \qquad N_{\nabla S_u} := \max_{\alpha \in B(\widehat{\alpha}, r) \cap \operatorname{dom} R} |||\nabla_\alpha S_u(\alpha)|||, \text{ and}$$

$$N_p := N_{\nabla S_u} + Cr \text{ with } C = L_{S_p} C_u.$$

(vii) The outer step length σ satisfies

$$0 < \sigma \leq \frac{1}{C_\alpha} \min\left\{ \frac{(C_F - 1)C_u}{L_{S_u} + C_F C_u}, \frac{(C_{F,S} - 1)C_p - (1 + C_{F,S})L_{S_p} C_u}{(1 + L_{S_u})L_{S_p} + C_{F,S} C_p - (1 + C_{F,S})L_{S_p} C_u} \right\}$$

with

$$C_F := \sqrt{1 + 2\tau\gamma_F(1-\kappa)}, \qquad C_{F,S} := \sqrt{(1+\theta\gamma_F)/(1-\theta\gamma_F)} \quad \text{and}$$
$$C_\alpha := N_p L_{\nabla J} C_u + N_{\nabla J} \max\{C_p, L_{S_p} C_u\} + L_\alpha + C_R.$$

(viii) The initial iterate (u^0, p^0, α^0) is such that the distance-to-solution

$$r_0 := \sqrt{\sigma\varphi_u \tau^{-1} \|u^0 - \widehat{u}\|^2 + \sigma\varphi_p \theta^{-1} |||p^0 - \widehat{p}|||^2 + \|\alpha^0 - \widehat{\alpha}\|^2} = \sqrt{\sigma}\|x^0 - \widehat{x}\|_{ZM}$$

satisfies

$$r_0 \leq r \quad \text{and} \quad r_0 \max\{2(C_u + L_{S_u}), \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}(1 + \tau L_F) + \tau L_{\nabla F,\widehat{u}}\} \leq r_u.$$

Remark 2.6 (Interpretation). The interpretation of Assumption 2.2 in Remark 2.4 also applies to Assumption 2.5. We stress that *to satisfy the inequality in (vi), it suffices to ensure small $C_u > 0$ by good relative initialization of u and p with respect to α, and choosing the testing parameters $\varphi_u, \varphi_p > 0$ small enough.* According to (viii), the latter can be done by initializing close to a solution, or if the radii $r_u > 0$ is large.

3 CONVERGENCE ANALYSIS

We now prove the convergence of the FEFB (Algorithm 2.1) and the FIFB (Algorithm 2.2) in the respective Sections 3.2 and 3.3. Before this we start with common results. Our proofs are self-contained, but follow on the "testing" approach of [49] (see also [12]). The main idea is to prove a monotonicity-type estimate for the operator H_{k+1} occurring in the implicit forms (2.8) and (2.12) of the algorithms,

and then use the three-point identity (1.2) with respect to ZM-norms and inner products. This yields an inequality that readily yields an estimate from which convergence rates can be observed. The main results for the FEFB and the FIFB and in the respective Theorems 3.16 and 3.21.

Throughout, we assume that either Assumption 2.2 (FEFB) or 2.5 (FIFB) holds, and *tacitly use the constants from the relevant one.* We also tacitly take it that $\alpha^k \in \operatorname{dom} R$ for all $k \in \mathbb{N}$, as this is guaranteed by the assumptions for $k = 0$, and by the proximal step in the algorithms for $k \geq 1$.

3.1 GENERAL RESULTS

Our main goal here is to bound the error in the inner and adjoint iterates u^k and p^k in terms of the outer iterates α^k. We also derive bounds on the outer steps, and local monotonicity estimates. We first show that the solution mapping for the adjoint equation (2.2) is Lipschitz.

Lemma 3.1. *Suppose $(u,\alpha) \mapsto \nabla_u^2 F(u;\alpha)$ and $(u,\alpha) \mapsto \nabla_{\alpha u} F(u;\alpha) \in P$ are Lipschitz continuous with the respective constants $L_{\nabla^2 F}$ and $L_{\nabla_{\alpha u} F}$ in some bounded closed set $V_u \times V_\alpha$. Also assume that $\gamma_F \cdot \mathrm{Id} \leq \nabla_u^2 F(u;\alpha)$ and $|||\nabla_{\alpha u} F||| \leq N_{\nabla_{\alpha u} F}$ in $V_u \times V_\alpha$ for some $\gamma_F, N_{\nabla_{\alpha u}} > 0$. Then S_p is Lipschitz continuous in $V_u \times V_\alpha$, i.e.*

$$|||S_p(u_1,\alpha_1) - S_p(u_2,\alpha_2)||| \leq L_{S_p}(\|u_1 - u_2\| + \|\alpha_1 - \alpha_2\|)$$

for $u_1, u_2 \in V_u$ and $\alpha_1, \alpha_2 \in V_\alpha$ with factor $L_{S_p} := \gamma_F^{-2} L_{\nabla^2 F} N_{\nabla_{\alpha u} F} + \gamma_F^{-1} L_{\nabla_{\alpha u} F}$.

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

$$\begin{aligned} S_p(u_1,\alpha_1) - S_p(u_2,\alpha_2) = {} & (\nabla_{\alpha u} F(u_1;\alpha_1) - \nabla_{\alpha u} F(u_2;\alpha_2))\,(\nabla_u^2 F(u_1;\alpha_1))^{-1} \\ & + \nabla_{\alpha u} F(u_2;\alpha_2)\left((\nabla_u^2 F(u_1;\alpha_1))^{-1} - (\nabla_u^2 F(u_2;\alpha_2))^{-1}\right). \end{aligned}$$

Thus the triangle inequality and the operator norm inequality Theorem B.1 (ii) give

$$\begin{aligned} & |||S_p(u_1,\alpha_1) - S_p(u_2,\alpha_2)||| \\ & \quad \leq \|(\nabla_u^2 F(u_1;\alpha_1))^{-1}\|\,|||\nabla_{\alpha u} F(u_1;\alpha_1) - \nabla_{\alpha u} F(u_2;\alpha_2)||| \\ & \qquad + |||\nabla_{\alpha u} F(u_2;\alpha_2)|||\,\|(\nabla_u^2 F(u_1;\alpha_1))^{-1} - (\nabla_u^2 F(u_2;\alpha_2))^{-1}\| =: E_1 + E_2. \end{aligned} \tag{3.1}$$

The assumption $\gamma_F \cdot \mathrm{Id} \leq \nabla_u^2 F(u;\alpha)$ implies $\|(\nabla_u^2 F(u;\alpha))^{-1}\| \leq \gamma_F^{-1}$. Therefore, also using the Lipschitz continuity of $(u,\alpha) \mapsto \nabla_{\alpha u} F(u;\alpha)$ in $V_u \times V_\alpha$, we get

$$E_1 \leq \gamma_F^{-1} L_{\nabla_{\alpha u} F}(\|u_1 - u_2\| + \|\alpha_1 - \alpha_2\|). \tag{3.2}$$

Towards estimating the second term on the right hand side of (3.1), 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. Then we use $|||\nabla_{\alpha u} F||| \leq N_{\nabla_{\alpha u} F}$ and the Lipschitz continuity of $\nabla_u^2 F(u;\alpha)$ to obtain

$$\begin{aligned} E_2 &= \|\nabla_u^2 F(u_1;\alpha_1)^{-1}(\nabla_u^2 F(u_1;\alpha_1) - \nabla_u^2 F(u_2;\alpha_2))\nabla_u^2 F(u_2;\alpha_2)^{-1}\| \cdot |||\nabla_{\alpha u} F(u_2;\alpha_2)||| \\ &\leq N_{\nabla_{\alpha u} F}\|\nabla_u^2 F(u_1;\alpha_1)^{-1}\|\,\|\nabla_u^2 F(u_2;\alpha_2)^{-1}\|\,\|(\nabla_u^2 F(u_1;\alpha_1) - \nabla_u^2 F(u_2;\alpha_2))\| \\ &\leq \gamma_F^{-2} L_{\nabla^2 F} N_{\nabla_{\alpha u} F}(\|u_1 - u_2\| + \|\alpha_1 - \alpha_2\|). \end{aligned}$$

Inserting this inequality and (3.2) into (3.1) establishes the claim. □

We now prove two simple step length bounds.

Lemma 3.2. *Let Assumption 2.2 or 2.5 hold. Then $\sigma < 1/L_\alpha$ and $1 < C_F < \sqrt{1 + \gamma_F / L_F}$.*

Proof. We have $C_F > 1$ since $\kappa < 1$ forces $2\tau\gamma_F(1-\kappa) > 0$. Assumption 2.2 (iv) or 2.5 (iv) implies $2\tau\gamma_F(1-\kappa) < 4\gamma_F(\kappa-\kappa^2)/L_F \leq \gamma_F/L_F$. Therefore $C_F < \sqrt{1+\gamma_F/L_F}$. For $C_F, C_u, L_{S_u} > 0$ it holds $C_F C_u - C_u < L_{S_u} + C_F C_u$. Hence Assumption 2.2 (vii) or 2.5 (vii) gives

$$\sigma \leq \frac{(C_F-1)C_u}{C_\alpha(L_{S_u}+C_F C_u)} < \frac{1}{C_\alpha} = \frac{1}{C_u(L_{S_p}N_{\nabla J}+N_p)+L_\alpha+C_R} < \frac{1}{L_\alpha}. \qquad \square$$

The next lemma explains the latter inequality for r_0 in Assumption 2.2 (viii) and 2.5 (viii). For u^n and α^n close enough to the respective solutions, it bounds the next iterate u^{n+1} and the true inner problem solution $S_u(\alpha^n)$ for α^n to the r_u-neighborhood of $\widehat{u}$.

Lemma 3.3. *Suppose Assumption 2.2 or 2.5 hold and $\alpha^n \in B(\widehat{\alpha}, r_0)$, as well as $u^n \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$. Then $u^{n+1} \in B(\widehat{u}, r_u)$ and $S_u(\alpha^n) \in B(\widehat{u}, r_u)$.*

Proof. The inner gradient step of Algorithm 2.1 or 2.2 with $u^n \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$ give

$$\|u^{n+1}-\widehat{u}\| \leq \|u^{n+1}-u^n\| + \|u^n-\widehat{u}\| \leq \tau\|\nabla_u F(u^n;\alpha^n)\| + \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0.$$

Using $\nabla_u F(\widehat{u};\widehat{\alpha}) = 0$, $\alpha^n \in B(\widehat{\alpha}, r_0)$, and the Lipschitz continuity of $F(\widehat{u};\cdot)$ and $F(\cdot;\alpha^n)$ from Assumption 2.2 (iii) or 2.5 (iii) we continue to estimate, as required

$$\begin{aligned}
\|u^{n+1}-\widehat{u}\| &\leq \tau\|\nabla_u F(u^n;\alpha^n) - \nabla_u F(\widehat{u};\alpha^n) + \nabla_u F(\widehat{u};\alpha^n) - \nabla_u F(\widehat{u};\widehat{\alpha})\| + \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0 \\
&\leq \tau(L_F\|u^n-\widehat{u}\| + L_{\nabla F,\widehat{u}}\|\alpha^n-\widehat{\alpha}\|) + \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0 \\
&\leq (\sqrt{\sigma^{-1}\varphi_u^{-1}\tau}(1+\tau L_F) + \tau L_{\nabla F,\widehat{u}})r_0 \leq r_u.
\end{aligned}$$

Next, the Lipschitz continuity of S_u in $B(\widehat{\alpha}, 2r)$ from Assumption 2.2 (ii) or 2.5 (ii) with $\alpha^n \in B(\widehat{\alpha}, r_0)$ and $r_0 \leq r$ from Assumption 2.2 (viii) or 2.5 (viii) imply

$$\|S_u(\alpha^n)-\widehat{u}\| = \|S_u(\alpha^n) - S_u(\widehat{\alpha})\| \leq L_{S_u}\|\alpha^n-\widehat{\alpha}\| \leq L_{S_u}r_0 \leq r_u. \qquad \square$$

We now introduce a working condition that we later prove. It guarantees that the Lipschitz and Hessian properties of Assumption 2.2(ii), (iii) and (v), or Assumption 2.5(ii), (iii) and (v) hold at iterates.

Assumption 3.4 (Iterate locality). Let $r_0 \leq r$ and N_p be defined in either Assumption 2.2 or 2.5. Then this assumption holds for a given $n \in \mathbb{N}$ if

$$\alpha^n \in B(\widehat{\alpha}, r_0), \quad u^n \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0), \quad \text{and} \quad |||p^{n+1}||| \leq N_p.$$

Indeed, the next two lemmas show that if Assumption 3.4 holds for $n = k$ along with some further conditions, then it holds for $n = k+1$.

Lemma 3.5. *Suppose either Assumption 2.2 or 2.5 holds. Let $n \in \mathbb{N}$ and suppose*

$$|||p^{n+1} - \nabla_\alpha S_u(\alpha^n)||| \leq C\|\alpha^n-\widehat{\alpha}\| \tag{3.3}$$

with $\alpha^n \in B(\widehat{\alpha}, r)$. Then $|||p^{n+1}||| \leq N_p$.

Proof. We estimate using (3.3) and the definitions of the relevant constants in Assumption 2.2 or 2.5 that

$$\begin{aligned}
|||p^{n+1}||| &\leq |||\nabla_\alpha S_u(\alpha^n)||| + |||p^{n+1} - \nabla_\alpha S_u(\alpha^n)||| \\
&\leq N_{\nabla S_u} + C\|\alpha^n-\widehat{\alpha}\| \leq N_{\nabla S_u} + Cr = N_p.
\end{aligned} \qquad \square$$

Lemma 3.6. *Let $k \in \mathbb{N}$. Suppose either Assumption 2.2 or 2.5 holds; Assumption 3.4 holds for $n = k$; and that (3.3) holds for $n = k+1$. If also $\|x^{n+1} - \widehat{x}\|_{ZM} \le \|x^n - \widehat{x}\|_{ZM}$ for $n \in \{0, \dots, k\}$, then Assumption 3.4 holds for $n = k+1$.*

Proof. Summing $\|x^{n+1} - \widehat{x}\|_{ZM} \le \|x^n - \widehat{x}\|_{ZM}$ over $n = 0, \dots, k$ gives $\|x^{k+1} - \widehat{x}\|_{ZM} \le \|x^0 - \widehat{x}\|_{ZM} = \sigma^{-1/2} r_0$. By the definitions of Z and M in (2.10) or (2.13), and (2.9b) or (2.12b) respectively, it follows $\alpha^{k+1} \in B(\widehat{\alpha}, r_0)$ and $u^{k+1} \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau} r_0)$ as required. We finish by using Lemma 3.5 with $n = k+1$ to establish $|||p^{k+2}||| \le N_p$. □

We next prove a monotonicity-type estimate for the inner objective. For this we need need the following three-point monotonicity inequality.

Theorem 3.7. *Let $z, \widehat{x} \in X$. Suppose $F \in C^2(X)$, and for some $L > 0$ and $\gamma \ge 0$ that $\gamma \cdot \mathrm{Id} \le \nabla^2 F(\zeta) \le L \cdot \mathrm{Id}$ for all $\zeta \in [\widehat{x}, z] := \{\widehat{x} + s(z - \widehat{x}) \mid s \in [0,1]\}$. Then*

$$\langle \nabla F(z) - \nabla F(\widehat{x}), x - \widehat{x} \rangle \ge \gamma(1-\beta)\|x - \widehat{x}\|^2 - \frac{L}{4\beta}\|x - z\|^2 \quad (\beta \in (0,1],\ x \in X). \tag{3.4}$$

Proof. The proof follows that of [12, Lemma 15.1] whose statement unnecessarily takes ζ in neighborhood of $\widehat{x}$ instead of just the interval $[\widehat{x}, z]$. □

Lemma 3.8. *Let $n \in \mathbb{N}$. Suppose either Assumption 2.2 or 2.5, and 3.4 hold. Then for any $\kappa \in (0,1)$, we have*

$$\begin{aligned}\langle \nabla_u F(u^n; \alpha^n), u^{n+1} - \widehat{u} \rangle \ge{}& \frac{\gamma_F(1-\kappa)}{2}\|u^{n+1} - \widehat{u}\|^2 \\ &- \frac{L_F}{4\kappa}\|u^{n+1} - u^n\|^2 - \frac{L^2_{\nabla F, \widehat{u}}}{2\gamma_F(1-\kappa)}\|\alpha^n - \widehat{\alpha}\|^2.\end{aligned} \tag{3.5}$$

Proof. Assumption 2.2 (iii) or 2.5 (iii) with $\alpha^n \in B(\widehat{\alpha}, r)$ and $u^n \in B(\widehat{u}, r_u)$ from Assumption 3.4 give $\gamma_F \cdot \mathrm{Id} \le \nabla_u^2 F(u; \alpha^n) \le L_F \cdot \mathrm{Id}$ for all $u \in [\widehat{u}, u^n]$. We have $\nabla_u F(\widehat{u}; \widehat{\alpha}) = 0$ since $0 \in H(\widehat{u}, \widehat{p}, \widehat{\alpha})$. Therefore Theorem 3.7 yields

$$\begin{aligned}\langle \nabla_u F(u^n; \alpha^n), u^{n+1} - \widehat{u} \rangle &= \langle \nabla_u F(u^n; \alpha^n) - \nabla_u F(\widehat{u}; \alpha^n) + \nabla_u F(\widehat{u}; \alpha^n) - \nabla_u F(\widehat{u}; \widehat{\alpha}), u^{n+1} - \widehat{u} \rangle \\ &\ge \gamma_F(1-\kappa)\|u^{n+1} - \widehat{u}\|^2 - \frac{L_F}{4\kappa}\|u^{n+1} - u^n\|^2 - |\langle \nabla_u F(\widehat{u}; \alpha^n) - \nabla_u F(\widehat{u}; \widehat{\alpha}), u^{n+1} - \widehat{u} \rangle|.\end{aligned}$$

Young's inequality and the definition of $L_{\nabla F, \widehat{u}}$ in Assumption 2.2 (iii) or 2.5 (iii) now readily establishes the claim. □

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

Lemma 3.9. *Let $n \in \mathbb{N}$. Suppose either Assumption 2.2 or 2.5 hold, as do Assumption 3.4, (3.3), and*

$$\|u^{n+1} - S_u(\alpha^n)\| \le C_u \|\alpha^n - \widehat{\alpha}\|. \tag{3.6}$$

Then

$$\|\alpha^{n+1} - \alpha^n\| \le \sigma[(N_p L_{\nabla J} C_u + N_{\nabla J} C + L_\alpha) + C_R]\|\alpha^n - \widehat{\alpha}\| \tag{3.7}$$

and

$$C_u\|\alpha^n - \widehat{\alpha}\| + L_{S_u}\|\alpha^{n+1} - \alpha^n\| \le C_F C_u\big(\|\alpha^n - \widehat{\alpha}\| - \|\alpha^{n+1} - \alpha^n\|\big) \le C_F C_u \|\alpha^{n+1} - \widehat{\alpha}\|. \tag{3.8}$$

Proof. Using the α-update of Algorithm 2.1 or 2.2, we estimate

$$\begin{aligned}\|\alpha^{n+1} - \alpha^n\| &= \|[\operatorname{prox}_{\sigma R}(\alpha^n - \sigma p^{n+1}\nabla_u J(u^{n+1})) - \alpha^n] - [\operatorname{prox}_{\sigma R}(\widehat{\alpha} - \sigma\widehat{p}\nabla_u J(\widehat{u})) - \widehat{\alpha}]\| \\ &\le \|[\operatorname{prox}_{\sigma R}(\alpha^n - \sigma p^{n+1}\nabla_u J(u^{n+1})) - \alpha^n] - [\operatorname{prox}_{\sigma R}(\alpha^n - \sigma\widehat{p}\nabla_u J(\widehat{u})) - \alpha^n]\| \\ &\quad + \|[\operatorname{prox}_{\sigma R}(\alpha^n - \sigma\widehat{p}\nabla_u J(\widehat{u})) - \alpha^n] - [\operatorname{prox}_{\sigma R}(\widehat{\alpha} - \sigma\widehat{p}\nabla_u J(\widehat{u})) - \widehat{\alpha}]\|.\end{aligned}$$

Since proximal maps are 1-Lipschitz, and R is by Assumption 2.2 (v) or 2.5 (v) locally prox-σ-contractive at $\widehat{\alpha}$ for $\widehat{p}\nabla_u J(\widehat{u})$ within $B(\widehat{\alpha}, r) \cap \operatorname{dom} R$ with factor C_R, it follows

$$\|\alpha^{n+1} - \alpha^n\| \le \sigma\|p^{n+1}\nabla_u J(u^{n+1}) - \widehat{p}\nabla_u J(\widehat{u})\| + \sigma C_R\|\alpha^n - \widehat{\alpha}\| =: \sigma Q + \sigma C_R\|\alpha^n - \widehat{\alpha}\|. \tag{3.9}$$

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 (v) or 2.5 (v). Hence

$$\begin{aligned}Q &\le \|p^{n+1}\nabla_u J(u^{n+1}) - \nabla_\alpha(J \circ S_u)(\alpha^n) + \nabla_\alpha(J \circ S_u)(\alpha^n) - \widehat{p}\,\nabla_u J(\widehat{u})\| \\ &\le \|p^{n+1}\nabla_u J(u^{n+1}) - \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n))\| + L_\alpha\|\alpha^n - \widehat{\alpha}\|.\end{aligned}$$

Using the Lipschitz continuity of $\nabla_u J$ from Assumption 2.2 (v) or 2.5 (v), we continue

$$\begin{aligned}Q &\le \|p^{n+1}(\nabla_u J(u^{n+1}) - \nabla_u J(S_u(\alpha^n)) + (p^{n+1} - \nabla_\alpha S_u(\alpha^n))\nabla_u J(S_u(\alpha^n))\| + L_\alpha\|\alpha^n - \widehat{\alpha}\| \\ &\le |||p^{n+1}|||L_{\nabla J}\|u^{n+1} - S_u(\alpha^n)\| + |||p^{n+1} - \nabla_\alpha S_u(\alpha^n)|||\|\nabla_u J(S_u(\alpha^n))\| + L_\alpha\|\alpha^n - \widehat{\alpha}\|.\end{aligned}$$

We have $|||p^{n+1}||| \le N_p$ and $\alpha^n \in B(\widehat{\alpha}, r)$ by Assumption 3.4. Hence $\|\nabla_u J(S_u(\alpha^n))\| \le N_{\nabla J}$ by the definition in Assumption 2.2 (vi) or 2.5 (vi). Using (3.3) and (3.6) therefore give

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

Inserting this into (3.9), we obtain (3.7). Assumption 2.2 (vii) or Assumption 2.5 (vii) and (3.7) then yield

$$(L_{S_u} + C_F C_u)\|\alpha^{n+1} - \alpha^n\| \le \sigma(L_{S_u} + C_F C_u)C_\alpha\|\alpha^n - \widehat{\alpha}\| \le (C_F - 1)C_u\|\alpha^n - \widehat{\alpha}\|.$$

Rearranging terms and finishing with the triangle inequality we get (3.8). □

Remark 3.10 (Gradient steps with respect to R). We could (in both FEFB and FIFB) also take a gradient step instead of a proximal step with respect to R with $L_{\nabla R}$-Lipschitz gradient. That is, we would perform for the outer problem the update

$$\alpha^{n+1} = \alpha^n - \sigma[p^{n+1}\nabla_u J(u^{n+1}) + \nabla R(\alpha^n)].$$

This can be shown to be convergent by changing (3.9) to

$$\begin{aligned}\|\alpha^{n+1} - \alpha^n\| &= \sigma\|p^{n+1}\nabla_u J(u^{n+1}) + \nabla R(\alpha^n)\| \\ &= \sigma\|p^{n+1}\nabla_u J(u^{n+1}) - \widehat{p}\,\nabla_u J(\widehat{u}) + \nabla R(\alpha^n) - \nabla R(\widehat{\alpha})\| \\ &\le \sigma\big(\|p^{n+1}\nabla_u J(u^{n+1}) - \widehat{p}\,\nabla_u J(\widehat{u})\| + L_{\nabla R}\|\alpha^n - \widehat{\alpha}\|\big).\end{aligned}$$

We next prove that if an inner problem iterate has small error, and we take a short step in the outer problem, then also the next inner problem iterate has small error.

Lemma 3.11. *Let $k \in \mathbb{N}$. Suppose Assumption 2.2 or 2.5 hold. If Assumption 3.4, (3.3), and (3.6) hold for $n = k$, then (3.6) holds for $n = k + 1$ and we have $\alpha^{k+1} \in B(\widehat{\alpha}, 2r_0)$.*

Proof. We plan to use Theorem 3.7 on $F(\cdot;\alpha^{k+1})$ followed by the three-point identity and simple manipulations. We begin by proving the conditions of the theorem.

First, we show that both $u^{k+1} \in B(\widehat{u}, r_u)$ and $S_u(\alpha^{k+1}) \in B(\widehat{u}, r_u)$. The former is immediate from Assumption 3.4 and Lemma 3.3. For the latter we use (3.8) of Lemma 3.9. Its first inequality readily implies either $\|\alpha^k - \widehat{\alpha}\| > \|\alpha^{k+1} - \alpha^k\|$ or $\alpha^{k+1} = \widehat{\alpha}$. In the latter case $S_u(\alpha^{k+1}) = \widehat{u} \in B(\widehat{u}, r_u)$. In the former, using $\alpha^k \in B(\widehat{\alpha}, r_0)$, we get

$$\|\alpha^{k+1} - \widehat{\alpha}\| \leq \|\alpha^{k+1} - \alpha^k\| + \|\alpha^k - \widehat{\alpha}\| < 2\|\alpha^k - \widehat{\alpha}\| \leq 2r_0$$

Therefore we can use the Lipschitz continuity of S_u in $B(\widehat{\alpha}, 2r)$ from Assumption 2.2 (ii) or 2.5 (ii) to estimate

$$\|S_u(\alpha^{k+1}) - \widehat{u}\| = \|S_u(\alpha^{k+1}) - S_u(\widehat{\alpha})\| \leq L_{S_u}\|\alpha^{k+1} - \widehat{\alpha}\| \leq L_{S_u} 2r_0.$$

This implies $S_u(\alpha^{k+1}) \in B(\widehat{u}, r_u)$ by Assumption 2.2 (viii) or 2.5 (viii).

Since both $u^{k+1}, S_u(\alpha^{k+1}) \in B(\widehat{u}, r_u)$, Assumption 2.2 (iii) or 2.5 (iii) shows that $\gamma_F \cdot \mathrm{Id} \leq \nabla^2 F(u) \leq L_F \cdot \mathrm{Id}$ for $u \in [S_u(\alpha^{k+1}), u^{k+1}]$. Consequently Theorem 3.7 and $\nabla_u F(S_u(\alpha^{k+1}); \alpha^{k+1}) = 0$ give

$$\langle \nabla_u F(u^{k+1};\alpha^{k+1}), u^{k+2} - S_u(\alpha^{k+1})\rangle \geq \gamma_F(1-\kappa)\|u^{k+2} - S_u(\alpha^{k+1})\|^2 - \frac{L_F}{4\kappa}\|u^{k+2} - u^{k+1}\|^2.$$

Inserting the u update of Algorithm 2.1 or 2.2, i.e., $-\tau^{-1}(u^{k+2} - u^{k+1}) = \nabla_u F(u^{k+1};\alpha^{k+1})$ and using the three-point identity (1.2) we get

$$\begin{aligned}\frac{1}{2\tau}\Big(\|u^{k+2} - S_u(\alpha^{k+1})\|^2 + \|u^{k+2} - u^{k+1}\|^2 - \|u^{k+1} - S_u(\alpha^{k+1})\|^2\Big)& \\ \leq -\gamma_F(1-\kappa)\|u^{k+2} - S_u(\alpha^{k+1})\|^2 + \frac{L_F}{4\kappa}\|u^{k+2} - u^{k+1}\|^2.&\end{aligned}$$

Equivalently

$$(1 + 2\tau\gamma_F(1-\kappa))\,\|u^{k+2} - S_u(\alpha^{k+1})\|^2 + \Big(1 - \frac{\tau L_F}{2\kappa}\Big)\|u^{k+2} - u^{k+1}\|^2 \leq \|u^{k+1} - S_u(\alpha^{k+1})\|^2.$$

Because Assumption 2.2 (iv) or 2.5 (iv) guarantees $1 - \tau L_F/(2\kappa) > 0$, this implies

$$\|u^{k+2} - S_u(\alpha^{k+1})\| \leq C_F^{-1}\|u^{k+1} - S_u(\alpha^{k+1})\|.$$

Therefore the triangle inequality, (3.6) for $n = k$ and the Lipschitz continuity of S_u in $B(\widehat{\alpha}, 2r) \ni \alpha^k, \alpha^{k+1}$ yield

$$\begin{aligned}\|u^{k+2} - S_u(\alpha^{k+1})\| \leq C_F^{-1}\|u^{k+1} - S_u(\alpha^{k+1})\| &\leq C_F^{-1}\big(\|u^{k+1} - S_u(\alpha^k)\| + L_{S_u}\|\alpha^{k+1} - \alpha^k\|\big)\\ &\leq C_F^{-1}\big(C_u\|\alpha^k - \widehat{\alpha}\| + L_{S_u}\|\alpha^{k+1} - \alpha^k\|\big).\end{aligned}$$

Inserting (3.8) here, we establish the claim. □

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 3.12. *Let $n \in \mathbb{N}$. Suppose Assumption 2.2 (v) and (vi), or 2.5 (v) and (vi) hold with Assumption 3.4 and*

$$\|u^{n+1} - S_u(\alpha^n)\| \leq C_u\|\alpha^n - \widehat{\alpha}\| \;\text{ and }\; \|\!|p^{n+1} - \nabla_\alpha S_u(\alpha^n)|\!\| \leq C\|\alpha^n - \widehat{\alpha}\|. \tag{3.10}$$

Then, for any $d > 0$,

$$\begin{aligned}\langle p^{n+1}\nabla_u J(u^{n+1}) + \partial R(\alpha^{n+1}), \alpha^{n+1} - \widehat{\alpha}\rangle &\geq -\frac{L_\alpha}{2}\|\alpha^{n+1} - \alpha^n\|^2\\ &+ \left(\frac{\gamma_\alpha}{2} - \frac{L_{\nabla J}N_p C_u + CN_{\nabla J}}{2d}\right)\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \left(\frac{\gamma_\alpha}{2} - \frac{(L_{\nabla J}N_p C_u + CN_{\nabla J})d}{2}\right)\|\alpha^n - \widehat{\alpha}\|^2.\end{aligned} \tag{3.11}$$

Proof. The α-update of both Algorithms 2.1 and 2.2 in implicit form reads

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

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 (3.11), these expressions and the monotonicity of ∂R yield

$$\begin{aligned} E_0 &= \langle p^{n+1}\nabla_u J(u^{n+1}) - \widehat{p}\,\nabla_u J(\widehat{u}) + q^{n+1} - \widehat{q}, \alpha^{n+1} - \widehat{\alpha}\rangle \\ &\geq \langle p^{n+1}\nabla_u J(u^{n+1}) - \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n)), \alpha^{n+1} - \widehat{\alpha}\rangle \\ &\quad + \langle \nabla_\alpha S_u(\alpha^n)\nabla_u J(S_u(\alpha^n)) - \widehat{p}\,\nabla_u J(\widehat{u}), \alpha^{n+1} - \widehat{\alpha}\rangle =: E_1 + E_2. \end{aligned} \tag{3.12}$$

We estimate E_1 and E_2 separately. The one-dimensional mean value theorem gives

$$E_2 = \langle \nabla_\alpha(J \circ S_u)(\alpha^n) - \nabla_\alpha(J \circ S_u)(\widehat{\alpha}), \alpha^{n+1} - \widehat{\alpha}\rangle = \langle Q(\alpha^n - \widehat{\alpha}), \alpha^{n+1} - \widehat{\alpha}\rangle$$

for some $\zeta \in [\widehat{\alpha}, \alpha^n]$ and $Q := \nabla_\alpha^2(J \circ S_u)(\zeta)$. Since $\|\alpha^n - \widehat{\alpha}\| \leq r$ by Assumption 3.4, also $\|\zeta - \widehat{\alpha}\| \leq r$. Therefore, the 3-point identity (1.2) and Assumption 2.2 (v) or 2.5 (v) yield

$$\begin{aligned} E_2 &= \frac{1}{2}\|\alpha^{n+1} - \widehat{\alpha}\|_Q^2 + \frac{1}{2}\|\alpha^n - \widehat{\alpha}\|_Q^2 - \frac{1}{2}\|\alpha^{n+1} - \alpha^n\|_Q^2 \\ &\geq \frac{\gamma_\alpha}{2}(\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \|\alpha^n - \widehat{\alpha}\|^2) - \frac{L_\alpha}{2}\|\alpha^{n+1} - \alpha^n\|^2. \end{aligned} \tag{3.13}$$

To estimate E_1 we rearrange

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

We have $\|\nabla_u J(S_u(\alpha^n))\| \leq N_{\nabla J}$ by the definition of the latter in Assumption 2.2 (vi) or 2.5 (vi) with $\alpha^n \in B(\widehat{\alpha}, r)$ from Assumption 3.4. The same assumptions establish that $\nabla_u J$ is Lipschitz. Hence, using the operator norm inequality Theorem B.1 (iii),

$$\begin{aligned} E_1 &\geq -|||p^{n+1}|||\,\|\nabla_u J(u^{n+1}) - \nabla_u J(S_u(\alpha^n))\|\,|||\alpha^{n+1} - \widehat{\alpha}\| \\ &\quad - |||p^{n+1} - \nabla_\alpha S_u(\alpha^n)|||\,\|\nabla_u J(S_u(\alpha^n))\|\,|||\alpha^{n+1} - \widehat{\alpha}\| \\ &\geq -\left(L_{\nabla J}|||p^{n+1}|||\,\|u^{n+1} - S_u(\alpha^n)\| + C\|\nabla_u J(S_u(\alpha^n))\|\,|||\alpha^n - \widehat{\alpha}\|\right)\|\alpha^{n+1} - \widehat{\alpha}\|. \end{aligned}$$

Applying (3.10) and Young's inequality now yields for any $d > 0$ the estimate

$$\begin{aligned} E_1 &\geq -\left(L_{\nabla J}N_p C_u + CN_{\nabla J}\right)\|\alpha^n - \widehat{\alpha}\|\,|||\alpha^{n+1} - \widehat{\alpha}\| \\ &\geq -\left(L_{\nabla J}N_p C_u + CN_{\nabla J}\right)\left(\frac{d}{2}\|\alpha^n - \widehat{\alpha}\|^2 + \frac{1}{2d}\|\alpha^{n+1} - \widehat{\alpha}\|^2\right). \end{aligned} \tag{3.14}$$

By inserting (3.13) and (3.14) into (3.12) we obtain the claim (3.11). □

3.2 CONVERGENCE: FORWARD–EXACT–FORWARD-BACKWARD

We now prove the convergence of Algorithm 2.1. We start with a lemma that shows an α-relative exactness estimate on the adjoint iterate when one holds for the inner iterate. This is needed to use Lemma 3.12. The main result of this subsection is in the final Theorem 3.16. It proves under Assumption 2.2 the linear convergence of $\{(u^n, \alpha^n)\}_{n\in\mathbb{N}}$ generated by Algorithm 2.1 to $(\widehat{u}, \widehat{\alpha})$ solving the first-order optimality condition (2.6) for some $\widehat{p}$.

Lemma 3.13. *Let $n \in \mathbb{N}$. Suppose Assumption 2.2 and the inner exactness estimate (3.6) hold as well as $\alpha^n \in B(\widehat{\alpha}, r_0)$ and $u^n \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$. Then (3.3) and (3.10) hold for $C = L_{S_p}C_u$.*

Proof. Since (3.3) with (3.6) equals (3.10), it suffices to prove (3.3). We assumed $\alpha^n \in B(\widehat{\alpha}, r_0)$ and $u^{n+1}, S_u(\alpha^n) \in B(\widehat{u}, r_u)$ by Lemma 3.3. Therefore the Lipschitz continuity of S_p in $B(\widehat{u}, r_u) \times B(\widehat{\alpha}, r)$ from Lemma 3.1 with Assumption 2.2 (ii) and (iii) and (3.6) give

$$|||p^{n+1} - \nabla_\alpha S_u(\alpha^n)||| = |||S_p(u^{n+1}, \alpha^n) - S_p(S_u(\alpha^n), \alpha^n)||| \le L_{S_p}\|u^{n+1} - S_u(\alpha^n)\| \le L_{S_p} C_u \|\alpha^n - \widehat{\alpha}\|. \quad \square$$

We are able to collect the previous lemmas into a descent estimate from which we immediately observe local linear convergence. We recall the definitions of the preconditioning and testing operators M and Z in (2.9b) and (2.10).

Lemma 3.14. *Let $n \in \mathbb{N}$ and suppose Assumptions 2.2 and 3.4, and the inner exactness estimate (3.6) hold. Then*

$$\|x^{n+1} - \widehat{x}\|_{ZM}^2 + 2\varepsilon_u\|u^{n+1} - \widehat{u}\|^2 + 2\varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2 \le \|x^n - \widehat{x}\|_{ZM}^2 \tag{3.15}$$

for $\varphi_u > 0$ as in Assumption 2.2 (vi),

$$\varepsilon_u := \frac{\varphi_u \gamma_F (1-\kappa)}{2} > 0, \quad \text{and} \quad \varepsilon_\alpha := \frac{\gamma_\alpha - (L_{\nabla J} N_p + L_{S_p} N_{\nabla J}) C_u}{2} > 0.$$

Proof. We start by proving the monotonicity estimate

$$\langle Z H_{n+1}(x^{n+1}), x^{n+1} - \widehat{x}\rangle \ge \mathscr{V}_{n+1}(\widehat{x}) - \frac{1}{2}\|x^{n+1} - x^n\|_{ZM}^2 \tag{3.16}$$

for $\mathscr{V}_{n+1}(\widehat{u}, \widehat{p}, \widehat{\alpha}) := \varepsilon_u\|u^{n+1} - \widehat{u}\|^2 + \varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2$. We observe that $\varepsilon_u, \varepsilon_\alpha > 0$ by Assumption 2.2. The monotonicity estimate (3.16) expands as

$$h_{n+1} \ge \mathscr{V}_{n+1}(\widehat{u}, \widehat{p}, \widehat{\alpha}) - \frac{\varphi_u}{2\tau}\|u^{n+1} - u^n\|^2 - \frac{1}{2\sigma}\|\alpha^{n+1} - \alpha^n\|^2 \tag{3.17}$$

for (all elements of the set)

$$h_{n+1} := \left\langle \begin{pmatrix} \varphi_u \nabla_u F(u^n; \alpha^n) \\ p^{n+1}\nabla_u^2 F(u^{n+1}; \alpha^n) + \nabla_{\alpha u} F(u^{n+1}; \alpha^n) \\ p^{n+1}\nabla_u J(u^{n+1}) + \partial R(\alpha^{n+1}) \end{pmatrix}, \begin{pmatrix} u^{n+1} - \widehat{u} \\ p^{n+1} - \widehat{p} \\ \alpha^{n+1} - \widehat{\alpha} \end{pmatrix} \right\rangle.$$

We estimate each of the three lines of h_{n+1} separately. For the first line, we use (3.5) from Lemma 3.8. For the middle line we observe that $p^{n+1}\nabla_u^2 F(u^{n+1}; \alpha^n) + \nabla_{\alpha u} F(u^{n+1}; \alpha^n) = 0$ by the p-update of Algorithm 2.1.

For the last line, we use (3.11) from Lemma 3.12 with $d = 2$. We can do this because (3.10) holds by (3.6) and Lemma 3.13. This gives

$$\langle p^{n+1}\nabla_u J(u^{n+1}) + \partial R(\alpha^{n+1}), \alpha^{n+1} - \widehat{\alpha}\rangle \ge -\frac{L_\alpha}{2}\|\alpha^{n+1} - \alpha^n\|^2 + \varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \varepsilon_\alpha\|\alpha^n - \widehat{\alpha}\|^2.$$

Summing with (3.5) we thus obtain

$$\begin{aligned} h_{n+1} \ge{}& \frac{\varphi_u \gamma_F (1-\kappa)}{2}\|u^{n+1} - \widehat{u}\|^2 - \frac{\varphi_u L_F}{4\kappa}\|u^{n+1} - u^n\|^2 - \frac{L_\alpha}{2}\|\alpha^{n+1} - \alpha^n\|^2 \\ &+ \varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2 + \left(\varepsilon_\alpha - \frac{\varphi_u L_{\nabla F, \widehat{u}}^2}{2\gamma_F(1-\kappa)}\right)\|\alpha^n - \widehat{\alpha}\|^2. \end{aligned}$$

The factor of the first term is ε_u and the factor of last term is zero. Since $\sigma < 1/L_\alpha$ by Lemma 3.2 and $L_F/(2\kappa) \le 1/\tau$ by Assumption 2.2 (iv), we obtain (3.17), i.e., (3.16).

We now come to the fundamental argument of the testing approach of [49], combining operator-relative monotonicity estimates with the three-point identity. Indeed, (3.16) combined with the implicit algorithm (2.8) gives

$$\langle ZM(x^{n+1} - x^n), x^{n+1} - \widehat{x}\rangle + \mathscr{V}_{n+1}(\widehat{x}) \leq \frac{1}{2}\|x^{n+1} - x^n\|^2_{ZM}.$$

Inserting the three-point identity (1.2) and expanding $\mathscr{V}_{n+1}$ yields (3.15). □

Before stating our main convergence result for the FEFB, we simplify the assumptions of the previous lemma to just Assumption 2.2.

Lemma 3.15. *Suppose Assumption 2.2 holds. Then* (3.15) *holds for any* $n \in \mathbb{N}$.

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

(*) *Assumption* 3.4, (3.6), and (3.15) hold.

We first prove (*) for $n = 0$. Assumption 2.2 (i) directly establishes (3.6). The definition of r_0 in Assumption 2.2 also establishes that $\alpha^n \in B(\widehat{\alpha}, r_0)$ and $u^n \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$. We have just proved the conditions of Lemma 3.13, which establishes (3.3) for $n = 0$. Now Lemma 3.5 establishes $|||p^1||| \leq N_p$. Therefore Assumption 3.4 holds for $n = 0$. Finally Lemma 3.14 proves (3.15) for $n = 0$. This concludes the proof of the induction base.

We then make the induction assumption that (*) holds for $n \in \{0, \dots, k\}$ and prove it for $n = k + 1$. Indeed, the induction assumption and Lemma 3.11 give (3.6) for $n = k + 1$. Next (3.15) for $n = k$ implies $\alpha^{k+1} \in B(\widehat{\alpha}, r_0)$ and $u^{k+1} \in B(\widehat{u}, \sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$, where r_0 and r_u are as in Assumption 2.2. Therefore Lemma 3.3 gives $u^{k+2} \in B(\widehat{u}, r_u)$ while Lemma 3.13 establishes (3.3) for $n = k + 1$. For all $n \in \{0, \dots, k\}$, the inequality (3.15) implies $\|x^{n+1} - \widehat{x}\|_{ZM} \leq \|x^n - \widehat{x}\|_{ZM}$. Therefore Lemma 3.6 proves Assumption 3.4 and finally Lemma 3.14 proves (3.15) and consequently (*) for $n = k + 1$. □

Theorem 3.16. *Suppose Assumption 2.2 holds. Then* $\varphi_u\tau^{-1}\|u^n - \widehat{u}\|^2 + \sigma^{-1}\|\alpha^n - \widehat{\alpha}\|^2 \longrightarrow 0$ *linearly.*

Proof. Lemma 3.15, expansion of (3.15), and basic manipulation show that

$$\begin{aligned}\varphi_u\tau^{-1}\|u^n - \widehat{u}\|^2 + \sigma^{-1}\|\alpha^n - \widehat{\alpha}\|^2 &\geq (\varphi_u\tau^{-1} + 2\varepsilon_u)\|u^{n+1} - \widehat{u}\|^2 + (\sigma^{-1} + 2\varepsilon_\alpha)\|\alpha^{n+1} - \widehat{\alpha}\|^2\\ &= (1 + 2\varepsilon_u\varphi_u^{-1}\tau)\varphi_u\tau^{-1}\|u^{n+1} - \widehat{u}\|^2 + (1 + 2\varepsilon_\alpha\sigma)\sigma^{-1}\|\alpha^{n+1} - \widehat{\alpha}\|^2\\ &\geq \mu\big(\varphi_u\tau^{-1}\|u^{n+1} - \widehat{u}\|^2 + \sigma^{-1}\|\alpha^{n+1} - \widehat{\alpha}\|^2\big)\end{aligned}$$

for $\mu := \min\{1 + 2\varepsilon_u\varphi_u^{-1}\tau, 1 + 2\varepsilon_\alpha\sigma\}$. Since $\mu > 1$, linear convergence follows. □

3.3 CONVERGENCE: FORWARD–INEXACT–FORWARD-BACKWARD

We now prove the convergence of Algorithm 2.2. The overall structure and idea of the proofs follows Section 3.2 and uses several lemmas from Section 3.1. We first prove monotonicity estimate lemma for the adjoint step and then that a small enough step length in the outer problem gurantees that the inner and adjoint iterates stay in a small local neighbourhood if they are already in one. The main result of this subsection is in the final Theorem 3.21. It proves under Assumption 2.5 the linear convergence of $\{(u^n, p^n, \alpha^n)\}_{n\in\mathbb{N}}$ generated by Algorithm 2.2 to $(\widehat{u}, \widehat{p}, \widehat{\alpha})$ solving the first-order optimality condition (2.6).

Lemma 3.17. *Let* $u \in U, \alpha \in \mathscr{A}$ *and* $p_1, p_2, \tilde{p} \in P$. *Moreover,* $\gamma_F \cdot \mathrm{Id} \leq \nabla^2_u F(u;\alpha) \leq L_F \cdot \mathrm{Id}$ *and*

$$\tilde{p}\nabla^2_u F(u;\alpha) + \nabla_{\alpha u} F(u;\alpha) = 0. \tag{3.18}$$

holds. Then

$$\langle\!\langle p_1\nabla^2_u F(u;\alpha) + \nabla_{\alpha u}(u;\alpha), p_2 - \tilde{p}\rangle\!\rangle \geq \frac{\gamma_F}{2}|||p_2 - \tilde{p}|||^2 + \frac{\gamma_F}{2}|||p_1 - \tilde{p}|||^2 - \frac{L_F}{2}|||p_2 - p_1|||^2.$$

Proof. Using (3.18), the three-point identity (1.2) and $\gamma_F \cdot \mathrm{Id} \le \nabla_u^2 F(u;\alpha) \le L_F \cdot \mathrm{Id}$ gives

$$\begin{aligned}
&\langle\!\langle p_1 \nabla_u^2 F(u;\alpha) + \nabla_{\alpha u} F(u;\alpha), p_2 - \tilde{p} \rangle\!\rangle \\
&= \langle\!\langle (p_1 - \tilde{p}) \nabla_u^2 F(u;\alpha), p_2 - \tilde{p} \rangle\!\rangle \\
&= \sum_{i \in I} \langle \nabla_u^2 F(u;\alpha)(p_1 - \tilde{p})^* \varphi_i, (p_2 - \tilde{p})^* \varphi_i \rangle \\
&= \sum_{i \in I} \left(\frac{1}{2} \|(p_1 - \tilde{p})^* \varphi_i\|^2_{\nabla_u^2 F(u;\alpha)} - \frac{1}{2} \|(p_2 - p_1)^* \varphi_i\|^2_{\nabla_u^2 F(u;\alpha)} + \frac{1}{2} \|(p_2 - \tilde{p})^* \varphi_i\|^2_{\nabla_u^2 F(u;\alpha)} \right) \\
&\ge \sum_{i \in I} \left(\frac{\gamma_F}{2} \|(p^{k+1} - \tilde{p})^* \varphi_i\|^2 - \frac{L_F}{2} \|(p^{k+2} - p^{k+1})^* \varphi_i\|^2 + \frac{\gamma_F}{2} \|(p^{k+2} - \tilde{p})^* \varphi_i\|^2 \right) \\
&= \frac{\gamma_F}{2} |||p_2 - \tilde{p}|||^2 + \frac{\gamma_F}{2} |||p_1 - \tilde{p}|||^2 - \frac{L_F}{2} |||p_2 - p_1|||^2.
\end{aligned}$$

□

Lemma 3.18. *Let $k \in \mathbb{N}$. Suppose Assumption 2.5 holds, and Assumption 3.4 and*

$$\|u^{n+1} - S_u(\alpha^n)\| \le C_u \|\alpha^n - \widehat{\alpha}\| \ \text{ and } \ |||p^{n+1} - \nabla_\alpha S_u(\alpha^n)||| \le C_p \|\alpha^n - \widehat{\alpha}\| \tag{3.19}$$

hold for $n = k$. Then (3.19) holds for $n = k + 1$.

Proof. Observe that (3.19) for $n = k$ implies (3.6) as well as (3.3) for $n = k$ and $C = C_p$. Lemma 3.11 therefore proves the first part of (3.19) for $n = k + 1$, i.e.,

$$\|u^{k+2} - S_u(\alpha^{k+1})\| \le C_u \|\alpha^{k+1} - \widehat{\alpha}\|. \tag{3.20}$$

We still need to show the second part $|||p^{k+2} - \nabla_\alpha S_u(\alpha^{k+1})||| \le C_p \|\alpha^{k+1} - \widehat{\alpha}\|$. We follow the fundamental argument of the testing approach (see the end of the proof of Lemma 3.14) and use Assumption 2.5 (ii) and (iii). For the latter we need $\alpha^k, \alpha^{k+1} \in B(\widehat{\alpha}, 2r)$ and $u^{k+2}, S_u(\alpha^k) \in B(\widehat{u}, r_u)$. We have $\alpha^k \in B(\widehat{\alpha}, r_0)$ by Assumption 3.4 and $\alpha^{k+1} \in B(\widehat{\alpha}, 2r_0)$ by Lemma 3.11. Thus we may use the Lipschitz continuity of S_u with the triangle inequality and (3.20) to get $S_u(\alpha^k) \in B(S_u(\widehat{\alpha}), L_{S_u} r_0) \subset B(\widehat{u}, r_u)$ and

$$\begin{aligned}
\|u^{k+2} - \widehat{u}\| &\le \|u^{k+2} - S_u(\alpha^{k+1})\| + \|S_u(\alpha^{k+1}) - S_u(\widehat{\alpha})\| \\
&\le (C_u + L_{S_u}) \|\alpha^{k+1} - \widehat{\alpha}\| \le (C_u + L_{S_u}) 2 r_0,
\end{aligned}$$

which yields $u^{k+2} \in B(\widehat{u}, r_u)$. The definition of S_p in (2.3) implies

$$S_p(u^{k+2}, \alpha^{k+1}) \nabla_u^2 F(u^{k+2}; \alpha^{k+1}) + \nabla_{\alpha u} F(u^{k+2}; \alpha^{k+1}) = 0.$$

Since also $\gamma_F \cdot \mathrm{Id} \le \nabla_u^2 F \le L_F \cdot \mathrm{Id}$ in $B(\widehat{u}, r_u) \times B(\widehat{\alpha}, 2r)$ from Assumption 2.5 (iii), we get

$$\begin{aligned}
&\langle\!\langle p^{k+1} \nabla_u^2 F(u^{k+2}; \alpha^{k+1}) + \nabla_{\alpha u} F(u^{k+2}; \alpha^{k+1}), p^{k+2} - S_p(u^{k+2}, \alpha^{k+1}) \rangle\!\rangle \\
&\ge \frac{\gamma_F}{2} |||p^{k+2} - S_p(u^{k+2}, \alpha^{k+1})|||^2 + \frac{\gamma_F}{2} |||p^{k+1} - S_p(u^{k+2}, \alpha^{k+1})|||^2 - \frac{L_F}{2} |||p^{k+2} - p^{k+1}|||^2 =: A
\end{aligned} \tag{3.21}$$

from Lemma 3.17. By the p update of the FIFB in the implicit form (2.11), we have

$$p^{k+1} \nabla_u^2 F(u^{k+2}; \alpha^{k+1}) + \nabla_{\alpha u} F(u^{k+2}; \alpha^{k+1}) = -\theta^{-1}(p^{k+2} - p^{k+1}).$$

Combining with (3.21) gives

$$-\langle\!\langle p^{k+2} - p^{k+1}, p^{k+2} - S_p(u^{k+2}, \alpha^{k+1}) \rangle\!\rangle \ge \theta A.$$

An application of the three-point identity (1.2) with $\theta L_F \le 1$ from Assumption 2.5 (iv) now yields for $C_{F,S} = \sqrt{(1+\theta\gamma_F)/(1-\theta\gamma_F)}$ the estimate

$$|||p^{k+2} - S_p(u^{k+2}, \alpha^{k+1})||| \le C_{F,S}^{-1} |||p^{k+1} - S_p(u^{k+2}, \alpha^{k+1})|||.$$

This estimate and the triangle inequality give

$$\begin{aligned} |||p^{k+2} - \nabla_\alpha S_u(\alpha^{k+1})||| &= |||p^{k+2} - S_p(S_u(\alpha^{k+1}), \alpha^{k+1})||| \\ &\le |||p^{k+2} - S_p(u^{k+2}, \alpha^{k+1})||| + |||S_p(u^{k+2}, \alpha^{k+1}) - S_p(S_u(\alpha^{k+1}), \alpha^{k+1})||| \\ &\le C_{F,S}^{-1} |||p^{k+1} - S_p(u^{k+2}, \alpha^{k+1})||| + |||S_p(u^{k+2}, \alpha^{k+1}) - S_p(S_u(\alpha^{k+1}), \alpha^{k+1})||| =: E_1 + E_2. \end{aligned} \tag{3.22}$$

The solution map S_u is Lipschitz in $B(\widehat{\alpha}, 2r)$ and S_p is Lipschitz in $B(\widehat{u}, r_u) \times B(\widehat{\alpha}, 2r)$ due to Assumption 2.5 (ii) and (iii), and Lemma 3.1. Combined with the triangle inequality, (3.19) for $n = k$ and (3.20), we obtain

$$\begin{aligned} C_{F,S} E_1 &\le |||p^{k+1} - S_p(S_u(\alpha^k), \alpha^k)||| + |||S_p(u^{k+2}, \alpha^{k+1}) - S_p(S_u(\alpha^k), \alpha^k)||| \\ &\le |||p^{k+1} - \nabla_\alpha S_u(\alpha^k))||| + L_{S_p} \left(\|u^{k+2} - S_u(\alpha^k)\| + \|\alpha^{k+1} - \alpha^k\| \right) \\ &\le C_p \|\alpha^k - \widehat{\alpha}\| + L_{S_p} \left(\|u^{k+2} - S_u(\alpha^{k+1})\| + \|S_u(\alpha^{k+1}) - S_u(\alpha^k)\| + \|\alpha^{k+1} - \alpha^k\| \right) \\ &\le E_3 + L_{S_p} C_u \|\alpha^{k+1} - \widehat{\alpha}\| \end{aligned} \tag{3.23}$$

for

$$E_3 := C_p \|\alpha^k - \widehat{\alpha}\| + L_{S_p}(1 + L_{S_u}) \|\alpha^{k+1} - \alpha^k\|.$$

Using again the Lipschitz continuity of S_p and (3.20), we get

$$E_2 \le L_{S_p} \|u^{k+2} - S_u(\alpha^{k+1})\| \le L_{S_p} C_u \|\alpha^{k+1} - \widehat{\alpha}\|. \tag{3.24}$$

Inserting (3.23) and (3.24) into (3.22) yields

$$|||p^{k+2} - \nabla_\alpha S_u(\alpha^{k+1})||| \le C_{F,S}^{-1} E_3 + (C_{F,S}^{-1} + 1) L_{S_p} C_u \|\alpha^{k+1} - \widehat{\alpha}\|.$$

Therefore the claim follows if we show that

$$C_{F,S}^{-1} E_3 \le (C_p - (C_{F,S}^{-1} + 1) L_{S_p} C_u) \|\alpha^{k+1} - \widehat{\alpha}\|. \tag{3.25}$$

Lemma 3.9 proves (3.7) with $C = C_p$. Together with Assumption 2.5 (vii) it yields

$$\|\alpha^{k+1} - \alpha^k\| \le \sigma C_\alpha \|\alpha^k - \widehat{\alpha}\| \le \frac{(C_{F,S} - 1) C_p - (1 + C_{F,S}) L_{S_p} C_u}{(1 + L_{S_u}) L_{S_p} + C_{F,S} C_p - (1 + C_{F,S}) L_{S_p} C_u} \|\alpha^k - \widehat{\alpha}\|.$$

Multiplying by $(1 + L_{S_u}) L_{S_p} + C_{F,S} C_p - (1 + C_{F,S}) L_{S_p} C_u$, rearranging terms, and continuing with the triangle inequality gives (3.25). Indeed,

$$\begin{aligned} E_3 &\le C_{F,S} (C_p - (C_{F,S}^{-1} + 1) L_{S_p} C_u) (\|\alpha^k - \widehat{\alpha}\| - \|\alpha^{k+1} - \alpha^k\|) \\ &\le C_{F,S} (C_p - (C_{F,S}^{-1} + 1) L_{S_p} C_u) \|\alpha^{k+1} - \widehat{\alpha}\|. \end{aligned}$$

□

We now show that the adjoint iterates stay local if the outer iterates do.

Again, by combining the previous lemmas, we prove an estimate from which local convergence is immediate. For this, we recall the definitions of the preconditioning and testing operators M and Z in (2.12b) and (2.10).

Lemma 3.19. *Suppose Assumptions 2.5 and 3.4, and the inner and adjoint exactness estimate (3.19) hold for $n \in \mathbb{N}$. Then*

$$\|x^{n+1} - \widehat{x}\|_{ZM}^2 + 2\varepsilon_u\|u^{n+1} - \widehat{u}\|^2 + 2\varepsilon_p\|p^n - \widehat{p}\|^2 + 2\varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2 \le \|x^n - \widehat{x}\|_{ZM}^2 \tag{3.26}$$

for

$$\varepsilon_u := \frac{\varphi_u\gamma_F(1-\kappa)}{2} - C_{\alpha,1} > 0, \quad \varepsilon_p := \frac{\varphi_p\gamma_F}{2} > 0, \quad \text{and} \quad \varepsilon_\alpha := \frac{\gamma_\alpha - C_{\alpha,1} - \sqrt{C_{\alpha,1}^2 + 4C_{\alpha,2}^2}}{2} > 0$$

where $\varphi_u, \varphi_p > 0$ are as in Assumption 2.5, $C_{\alpha,1} := \varphi_p\frac{L_F L_{S_p}}{\gamma_F}$, and $C_{\alpha,2} := \left(L_{\nabla J}N_p + L_{S_p}N_{\nabla J}\right)\frac{C_u}{2}$.

Proof. We start by proving the monotonicity estimate

$$\langle ZH_{n+1}(x^{n+1}), x^{n+1} - \widehat{x}\rangle \ge \mathscr{V}_{n+1}(\widehat{x}) - \frac{1}{2}\|x^{n+1} - x^n\|_{ZM}^2 \tag{3.27}$$

for $\mathscr{V}_{n+1}(\widehat{u},\widehat{p},\widehat{\alpha}) = \varepsilon_u\|u^{n+1} - \widehat{u}\|^2 + \varepsilon_p\|p^n - \widehat{p}\|^2 + \varepsilon_\alpha\|\alpha^{n+1} - \widehat{\alpha}\|^2$. We observe that $\varepsilon_u, \varepsilon_p, \varepsilon_\alpha > 0$ by Assumption 2.5. The monotonicity estimate (3.27) expands as

$$h_{n+1} \ge \mathscr{V}_{n+1}(\widehat{u},\widehat{p},\widehat{\alpha}) - \frac{\varphi_u}{2\tau}\|u^{n+1} - u^n\|^2 - \frac{\varphi_p}{2\theta}|||p^{n+1} - p^n|||^2 - \frac{1}{2\sigma}\|\alpha^{n+1} - \alpha^n\|^2 \tag{3.28}$$

for (all elements of the set)

$$h_{n+1} := \left\langle \begin{pmatrix} \varphi_u\nabla_u F(u^n;\alpha^n) \\ \varphi_p\left(p^n\nabla_u^2 F(u^{n+1};\alpha^n) + \nabla_{\alpha u}F(u^{n+1};\alpha^n)\right) \\ p^{n+1}\nabla_u J(u^{n+1}) + \partial R(\alpha^{n+1}) \end{pmatrix}, \begin{pmatrix} u^{n+1} - \widehat{u} \\ p^{n+1} - \widehat{p} \\ \alpha^{n+1} - \widehat{\alpha} \end{pmatrix} \right\rangle.$$

We estimate the three lines of h_{n+1} separately. We immediately take care of the first line by using (3.5) from Lemma 3.8.

For the second line, using the optimality condition (2.2) we have

$$\begin{aligned} &\varphi_p\langle\!\langle p^n\nabla_u^2 F(u^{n+1};\alpha^n) + \nabla_{\alpha u}F(u^{n+1};\alpha^n), p^{n+1} - \widehat{p}\rangle\!\rangle \\ &= \varphi_p\langle\!\langle (p^n - S_p(u^{n+1},\alpha^n))\nabla_u^2 F(u^{n+1};\alpha^n), p^{n+1} - \widehat{p}\rangle\!\rangle \\ &= \varphi_p\langle\!\langle (p^n - \widehat{p})\nabla_u^2 F(u^{n+1};\alpha^n), p^{n+1} - \widehat{p}\rangle\!\rangle + \varphi_p\langle\!\langle (\widehat{p} - S_p(u^{n+1},\alpha^n))\nabla_u^2 F(u^{n+1};\alpha^n), p^{n+1} - \widehat{p}\rangle\!\rangle \\ &=: \varphi_p(E_1 + E_2). \end{aligned} \tag{3.29}$$

We have $u^{n+1}, S_u(\alpha^n) \in B(\widehat{u}, r_u)$, and $\alpha^n \in B(\widehat{\alpha}, r)$ by Lemma 3.3 and Assumption 3.4. Thus $\gamma_F \cdot \mathrm{Id} \le \nabla_u^2 F \le L_F \cdot \mathrm{Id}$ in $B(\widehat{u}, r_u) \times B(\widehat{\alpha}, 2r)$ and $\|\nabla_u^2 F(u^{n+1};\alpha^n)\| \le L_F$ by Assumption 2.5 (iii). We get

$$E_1 \ge \frac{\gamma_F}{2}|||p^{n+1} - \widehat{p}|||^2 + \frac{\gamma_F}{2}|||p^n - \widehat{p}|||^2 - \frac{L_F}{2}|||p^{n+1} - p^n|||^2 \tag{3.30}$$

from Lemma 3.17. By Theorem B.1 (i) $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is an inner product and $|||\cdot|||$ a norm on $\mathbb{L}(U;\mathscr{A})$, and we can use thus Cauchy–Schwarz inequality for them. Therefore, using also Theorem B.1 (ii), Lemma 3.1 and Young's inequality, we can estimate

$$\begin{aligned} E_2 &\ge -\left|\langle\!\langle (\widehat{p} - S_p(u^{n+1},\alpha^n))\nabla_u^2 F(u^{n+1};\alpha^n), p^{n+1} - \widehat{p}\rangle\!\rangle\right| \\ &\ge -|||(\widehat{p} - S_p(u^{n+1},\alpha^n))\nabla_u^2 F(u^{n+1};\alpha^n)|||\,|||p^{n+1} - \widehat{p}||| \\ &\ge -L_F|||S_p(\widehat{u},\widehat{\alpha}) - S_p(u^{n+1},\alpha^n))|||\,|||p^{n+1} - \widehat{p}||| \\ &\ge -L_F L_{S_p}\left(\|u^{n+1} - \widehat{u}\| + \|\alpha^n - \widehat{\alpha}\|\right)|||p^{n+1} - \widehat{p}||| \\ &\ge -\frac{L_F L_{S_p}}{\gamma_F}\left(\|u^{n+1} - \widehat{u}\|^2 + \|\alpha^n - \widehat{\alpha}\|^2\right) - \frac{\gamma_F}{2}|||p^{n+1} - \widehat{p}|||^2. \end{aligned} \tag{3.31}$$

Inserting (3.30) and (3.31) into (3.29), we obtain

$$\begin{aligned}(3.32)\qquad &\varphi_p\langle\!\langle p^n\nabla_u^2 F(u^{n+1};\alpha^n)+\nabla_{\alpha u}F(u^{n+1};\alpha^n),p^{n+1}-\widehat{p}\rangle\!\rangle\\ &\geq \frac{\varphi_p\gamma_F}{2}|||p^n-\widehat{p}|||^2-\frac{\varphi_p L_F}{2}|||p^{n+1}-p^n|||^2-\frac{\varphi_p L_F L_{S_p}}{\gamma_F}\left(\|u^{n+1}-\widehat{u}\|^2+\|\alpha^n-\widehat{\alpha}\|^2\right)\end{aligned}$$

The factor of the last term equals $C_{\alpha,1}$ from Assumption 2.5 (v).

Since Assumption 3.4 and (3.19) hold, Lemma 3.12 gives (3.11) with $C = C_p$ and any $d > 0$ for the third line of h_{n+1}. Summing (3.5), (3.11) and (3.32) we finally deduce

$$\begin{aligned}h_{n+1}\geq{}&\left(\frac{\varphi_u\gamma_F(1-\kappa)}{2}-C_{\alpha,1}\right)\|u^{n+1}-\widehat{u}\|^2-\frac{\varphi_u L_F}{4\kappa}\|u^{n+1}-u^n\|^2+\frac{\varphi_p\gamma_F}{2}|||p^n-\widehat{p}|||^2-\frac{\varphi_p L_F}{2}|||p^{n+1}-p^n|||^2\\ &-\frac{L_\alpha}{2}\|\alpha^{n+1}-\alpha^n\|^2+\left(\frac{\gamma_\alpha}{2}-\frac{C_{\alpha,2}}{d}\right)\|\alpha^{n+1}-\widehat{\alpha}\|^2+\left(\frac{\gamma_\alpha}{2}-\frac{\varphi_u L^2_{\nabla F,\widehat{u}}}{2\gamma_F(1-\kappa)}-C_{\alpha,1}-C_{\alpha,2}d\right)\|\alpha^n-\widehat{\alpha}\|^2.\end{aligned}$$

We have

$$\frac{C_{\alpha,2}}{d}=C_{\alpha,1}+C_{\alpha,2}d=\frac{C_{\alpha,1}}{2}+\frac{\sqrt{C_{\alpha,1}^2+4C_{\alpha,2}^2}}{2}\quad\text{for}\quad d=\frac{-C_{\alpha,1}+\sqrt{C_{\alpha,1}^2+4C_{\alpha,2}^2}}{2C_{\alpha,2}}.$$

Then also $\frac{\gamma_\alpha}{2}-\frac{C_{\alpha,2}}{d}=\varepsilon_\alpha$. It follows

$$h_{n+1}\geq\varepsilon_u\|u^{n+1}-\widehat{u}\|^2-\frac{\varphi_u L_F}{4\kappa}\|u^{n+1}-u^n\|^2+\varepsilon_p\|p^n-\widehat{p}\|^2-\frac{\varphi_p L_F}{2}|||p^{n+1}-p^n|||^2-\frac{L_\alpha}{2}\|\alpha^{n+1}-\alpha^n\|^2+\varepsilon_\alpha\|\alpha^{n+1}-\widehat{\alpha}\|^2.$$

Since $\sigma<1/L_\alpha$ by Lemma 3.2, $L_F/(2\kappa)\leq 1/\tau$ and $\theta<1/L_F$ by Assumption 2.5 (iv), we obtain (3.28), i.e., (3.27). We finish by applying the fundamental arguments of the testing approach to (3.27) and the general implicit update (2.8) as in Lemma 3.14. □

We simplify the assumptions of the previous lemma to just Assumption 2.5.

Lemma 3.20. *Suppose Assumption 2.5 holds. Then (3.26) holds for any $n\in\mathbb{N}$.*

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

(**) *Assumption* 3.4, (3.19), and (3.26) hold.

We first prove (**) for $n=0$. Assumption 2.5 (i) directly establishes (3.19). The definition of r_0 in Assumption 2.5 also establishes that $\alpha^n\in B(\widehat{\alpha},r_0)$ and $u^n\in B(\widehat{u},\sqrt{\sigma^{-1}\varphi_u^{-1}\tau}r_0)$. We have just proved the conditions of Lemma 3.5, which gives $|||p^1|||\leq N_p$. Thus we we proved Assumption 3.4 for $n=0$. Now Lemma 3.19 proves (3.26) for $n=0$. This concludes the proof of induction base.

We then make the induction assumption that (**) holds for $n\in\{0,\dots,k\}$ and prove it for $n=k+1$. The induction assumption and Lemma 3.18 give (3.19) for $n=k+1$. The inequality (3.26) for $n\in\{0,\dots,k\}$ also ensures $\|x^{n+1}-\widehat{x}\|_{ZM}\leq\|x^n-\widehat{x}\|_{ZM}$ for $n\in\{0,\dots,k\}$. Therefore Lemma 3.6 proves Assumption 3.4 for $n=k+1$. Now Lemma 3.19 shows (3.26) and concludes the proof of (**) for $n=k+1$. □

We finally come to the main convergence result for the FIFB.

Theorem 3.21. *Suppose Assumption 2.5 holds. Then $\varphi_u\tau^{-1}\|u^n-\widehat{u}\|^2+\varphi_p\theta^{-1}\|p^n-\widehat{p}\|^2+\sigma^{-1}\|\alpha^n-\widehat{\alpha}\|^2\to 0$ linearly.*

Proof. We define $\mu_1 := \min\{(1+2\varepsilon_u\varphi_u^{-1}\tau), (1+2\varepsilon_\alpha\sigma)\}$ and $\mu_2 := 1-2\varepsilon_p\varphi_p^{-1}\theta$. Lemma 3.20, expansion of (3.26), and basic manipulation show that

$$
\begin{aligned}
(3.33)\qquad & \mu_1\big(\varphi_u\tau^{-1}\|u^{n+1}-\widehat{u}\|^2+\sigma^{-1}\|\alpha^{n+1}-\widehat{\alpha}\|^2\big)+\varphi_p\theta^{-1}|||p^{n+1}-\widehat{p}|||^2\\
&\le (1+2\varepsilon_u\varphi_u^{-1}\tau)\varphi_u\tau^{-1}\|u^{n+1}-\widehat{u}\|^2+(1+2\varepsilon_\alpha\sigma)\sigma^{-1}\|\alpha^{n+1}-\widehat{\alpha}\|^2+\varphi_p\theta^{-1}|||p^{n+1}-\widehat{p}|||^2\\
&= (\varphi_u\tau^{-1}+2\varepsilon_u)\|u^{n+1}-\widehat{u}\|^2+(\sigma^{-1}+2\varepsilon_\alpha)\|\alpha^{n+1}-\widehat{\alpha}\|^2+\varphi_p\theta^{-1}|||p^{n+1}-\widehat{p}|||^2\\
&\le \varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\mu_2\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2.
\end{aligned}
$$

There are two separate cases (i) $\mu_1\mu_2 \le 1$ and (ii) $\mu_1\mu_2 > 1$. In case (i), we have

$$
\begin{aligned}
&\varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\mu_2\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2\\
&= \mu_1^{-1}\big(\mu_1\big(\varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2\big)+\mu_1\mu_2\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2\big)\\
&\le \mu_1^{-1}\big(\mu_1\big(\varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2\big)+\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2\big),
\end{aligned}
$$

which with (3.33) implies linear convergence since $\mu_1^{-1} \in (0,1)$. In case (ii), we obtain

$$
\begin{aligned}
&\varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\mu_2\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2\\
&\le \mu_2\big(\mu_1\big(\varphi_u\tau^{-1}\|u^{n}-\widehat{u}\|^2+\sigma^{-1}\|\alpha^{n}-\widehat{\alpha}\|^2\big)+\varphi_p\theta^{-1}|||p^{n}-\widehat{p}|||^2\big),
\end{aligned}
$$

which with (3.33) implies linear convergence since $\mu_2 \in (0,1)$. □

4 NUMERICAL EXPERIMENTS

We evaluate the performance of our proposed algorithms on parameter learning for (anisotropic, smoothed) total variation image denoising and deconvolution. For a "ground truth" image $b \in \mathbb{R}^{N^2}$ of dimensions $N \times N$, we take

$$J(u) = \frac{1}{2}\|u-b\|_2^2$$

as the outer fitness function. For b we use a cropped portion of image 02 or 08 from the free Kodak dataset [26] converted to gray values in $[0,1]$. The purpose of these numerical experiments is a simple performance comparison between our proposed methods and a few representative approaches from the literature. We therefore only consider a single ground-truth image b and a corresponding corrupted data z in the various inner problems, which we next describe. For proper generalizable parameter learning, multiple such training pairs (b_i, z_i) should be used. This can in principle be done by summing over all the data in both the inner and outer problem; resulting in a higher-dimensional bilevel problem; see, e.g., [5]. In practise, a large sample count would require stochastic techniques.

4.1 DENOISING

For denoising we take in (1.1) as the inner objective

$$F(u;\alpha) = \frac{1}{2}\|u-z\|_2^2 + \alpha\|Du\|_{1,\gamma} \quad (u \in \mathbb{R}^{N^2}, \alpha \in [0,\infty)),$$

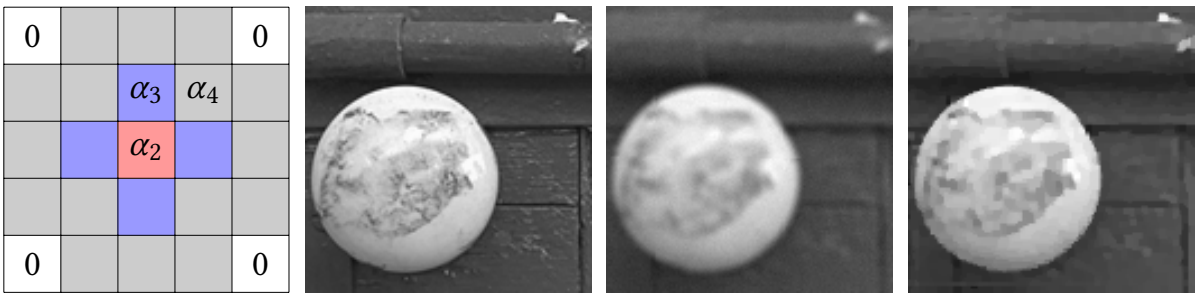
and as the outer regulariser $R \equiv 0$. The simulated measurement z is obtained from b by adding Gaussian noise of standard deviation 0.1. The matrix D is a backward difference operator with Dirichlet boundary conditions. Instead of the one-norm, $\|\cdot\|_1$, to ensure the twice differentiability of the objective and hence a simple adjoint equation, we use a C^2 Huber- or Moreau–Yosida -type approximation with

$$\|y\|_{1,\gamma} := \sum_{i=1}^{2N^2} \rho_\gamma(y_i), \quad \text{where} \quad \rho_\gamma(x) := \begin{cases} -\frac{|x|^3}{3\gamma^2}+\frac{|x|^2}{\gamma} & \text{if } |x| \le \gamma,\\ |x|-\frac{\gamma}{3} & \text{if } |x| > \gamma. \end{cases}$$

We used $\gamma = 10^{-4}$ in our experiments.

(a) Original (b) Noisy; error 14.1% (c) Implicit; error 9.3% (d) FIFB; error 9.2%

Figure 1: Denoising data and results for the implicit and FIFB methods ($N = 128$).



(a) Kernel structure (b) Original (c) Blurry; error 9.8% (d) FIFB; error 7.2%

Figure 2: Deconvolution kernel parametrisation, data, and result for FIFB ($N = 128$).

4.2 DECONVOLUTION

For deconvolution, we take as the inner objective

$$F(u;\alpha) = \frac{1}{2}\|K(\alpha) * u - z\|_2^2 + C\alpha_1\|Du\|_{1,\gamma}, \quad (u \in \mathbb{R}^{N^2}, \alpha \in [0,\infty)^4),$$

and as the outer regulariser $R(\alpha) = \beta(\sum_{i=2}^4 \alpha_i - 1)^2 + \delta_{[0,\infty)}(\alpha_1)$ for a regularisation parameter $\beta = 10^4$. We introduce the constant $C = \frac{1}{10}$ to help convergence by ensuring the same order of magnitude for all components of α. The first element of α is the total variation regularization parameter while the rest parametrizes the convolution kernel $K(\alpha)$ as illustrated in Figure 2a. The sum of the elements of the kernel equals $\alpha_2+\alpha_3+\alpha_4$. Operator r_θ rotates image θ degrees, clockwise for $\theta > 0$ and counterclocwise for $\theta < 0$. We form z by computing $r_{-1}(K(\alpha) * r_1(b))$ for $\alpha = [0.15, 0.1, 0.75]$ and adding Gaussian noise of standard deviation $1 \cdot 10^{-2}$.

For denoising, and deconvolution assuming $\ker D \cap \ker K(\alpha) = \{0\}$, it is not difficult to verify the structural parts of Assumptions 2.2 and 2.5, required for the convergence results of Section 3. We do not attempt to verify the conditions on the step lengths, choosing them by trial and error.

4.3 AN IMPLICIT BASELINE METHOD

We evaluate Algorithms 2.1 and 2.2 against a conventional method that solves both the inner problem and the adjoint equation (near-)exactly, as well as the AID [35]. We also experimented with solving the equivalent constrained optimisation problem $\min_{\alpha,u} J(u)$ subject to $\nabla_u F(u;\alpha) = 0$ with IPOPT [52] and the NL-PDPS [11, 48]. However, we did not observe convergence without the inclusion of additional H^1 regularisation in the inner problem, as in, e.g., [14]. Since that changes the problem, we have not included "whole problem" approaches in our comparison.

Table 1: Algorithm parametrisation (step length parameters, inner and adjoint iteration counts), time multiplier, and outer steps taken to reach threshold computational resources (CPU time) value. The threshold is 6000 for denoising, and 15000 for deconvolution. The time multipliers allows conversion from computational resources to real time. It differs between algorithms and problem dimensions due to different levels of parallelisability. The 'inner steps' and 'adjoint steps' columns indicate the (maximum) number of iterations taken towards a solution of the inner problem or the adjoint equation on every outer iteration. The `bicgstab` method for solving the adjoint for the implicit method and the FEFB may use a smaller number of iterations as determined by the threshold 10^{-5}. "LS" for the step length parameter σ means that line search was used.

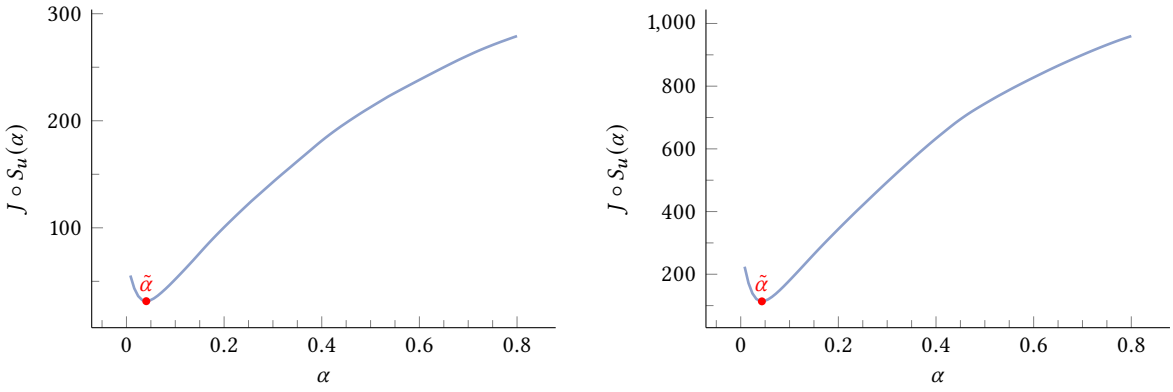
	N	method	outer steps	inner steps	adjoint steps	time mult.	τ	θ	σ
denoising	128	implicit	15	$2.5 \cdot 10^4$	10^3	0.22	$5 \cdot 10^{-4}$	-	LS
		AID	$7.7 \cdot 10^3$	10	50	0.12	$5 \cdot 10^{-4}$	-	$2 \cdot 10^{-8}$
		FEFB	$3.9 \cdot 10^3$	1	10^3	0.11	$8 \cdot 10^{-4}$	-	$5 \cdot 10^{-9}$
		FIFB	$5.1 \cdot 10^5$	1	1	0.22	$5 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$5 \cdot 10^{-10}$
	256	implicit	12	$2.5 \cdot 10^4$	10^3	0.40	$5 \cdot 10^{-4}$	-	LS
		AID	$3.9 \cdot 10^3$	10	50	0.21	$5 \cdot 10^{-4}$	-	$1 \cdot 10^{-8}$
		FEFB	$2.3 \cdot 10^3$	1	10^3	0.16	$8 \cdot 10^{-4}$	-	$2 \cdot 10^{-9}$
		FIFB	$2.3 \cdot 10^5$	1	1	0.35	$5 \cdot 10^{-4}$	$7.5 \cdot 10^{-6}$	$5 \cdot 10^{-10}$
deconvolution	128	implicit	29	$1 \cdot 10^5$	10^3	0.21	$5 \cdot 10^{-4}$	-	$5 \cdot 10^{-5}$
		AID	$11 \cdot 10^3$	50	50	0.14	$5 \cdot 10^{-4}$	-	$2.5 \cdot 10^{-6}$
		FEFB	$4.9 \cdot 10^2$	1	10^3	0.15	$5 \cdot 10^{-4}$	-	$1 \cdot 10^{-6}$
		FIFB	$5.2 \cdot 10^5$	1	1	0.22	$5 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$2.5 \cdot 10^{-7}$
	32	implicit	$8.7 \cdot 10^2$	$5 \cdot 10^4$	10^3	0.93	$5 \cdot 10^{-4}$	-	$5 \cdot 10^{-6}$
		AID	$9.2 \cdot 10^5$	10	50	1.00	$5 \cdot 10^{-4}$	-	$1 \cdot 10^{-7}$
		FEFB	$11 \cdot 10^3$	1	10^3	0.19	$5 \cdot 10^{-4}$	-	$1 \cdot 10^{-6}$
		FIFB	$4.4 \cdot 10^6$	1	1	0.65	$5 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$7 \cdot 10^{-8}$

To solve the inner problem in the implicit baseline method, we use gradient descent, starting at $v^0 = 0$ and updating $v^{m+1} := v^m - \tau_m \nabla F(v^m; \alpha^k)$ We then set $u^{k+1} = v^{m+1}$. The adjoint and outer iterate updates are as in Algorithm 2.1, however, we discover $\sigma = \sigma_k$ by the line search rule [12, (12.41)] for nonsmooth problems, starting at $\sigma_k = 5 \cdot 10^{-5}$ and multiplying by 0.1 on each line search step. For deconvolution we use a fixed step length parameter, as it performed better. The specific parameter choices (step lengths, number of inner and adjoint iterations) for all algorithms and experiments are listed in Table 1.

4.4 NUMERICAL SETUP

Our algorithm implementations are available on Zenodo [47]. To solve the adjoint equation in the FEFB and implicit methods, we use Matlab's `bicgstab` implementation of the stabilized biconjugate gradient method [51] with tolerance 10^{-5}, and maximum iteration count 10^3. With the AID we use 50 conjugate gradient iterations. These choices, as well as the choice of the number of inner iterations for the implicit method and the AID, have been made by trial and error to be as small as possible while obtaining an apparently stable algorithm.

To evaluate scalability, we consider for denoising both $N = 128$ and $N = 256$. For deconvolution we consider $N = 128$ and $N = 32$. We take initial $u^0 = S_u(\alpha^0)$ and $p^0 = S_p(u^0, \alpha^0)$ where for denoising



(a) $N = 128$ (b) $N = 256$

Figure 3: Graph of composed objective the denoising problem (both problem sizes), along with near-optimal $\tilde{\alpha}$ found by recursive subdivision.

$\alpha^0 = 0$ and for deconvolution $\alpha^0 = [0.4, 0.25, 0.25, 0.5]$ and $\alpha^0 = [0.04, 0.25, 0.25, 0.5]$ with $N = 128$ and $N = 32$ respectively.

To compare algorithm performance, we plot relative errors against the `cputime` value of Matlab on an AMD Ryzen 5 5600H CPU. We call this value "computational resources", as it takes into account the use of several CPU cores by Matlab's internal linear algebra, making it fairer than the actual running time. For each algorithm and problem, we indicate in Table 1 the step length parameters, the number of outer steps to reach the computational resources value of 6000 for denoising 15000 for deconvolution , and an average multiplier to convert computational resources into running times.

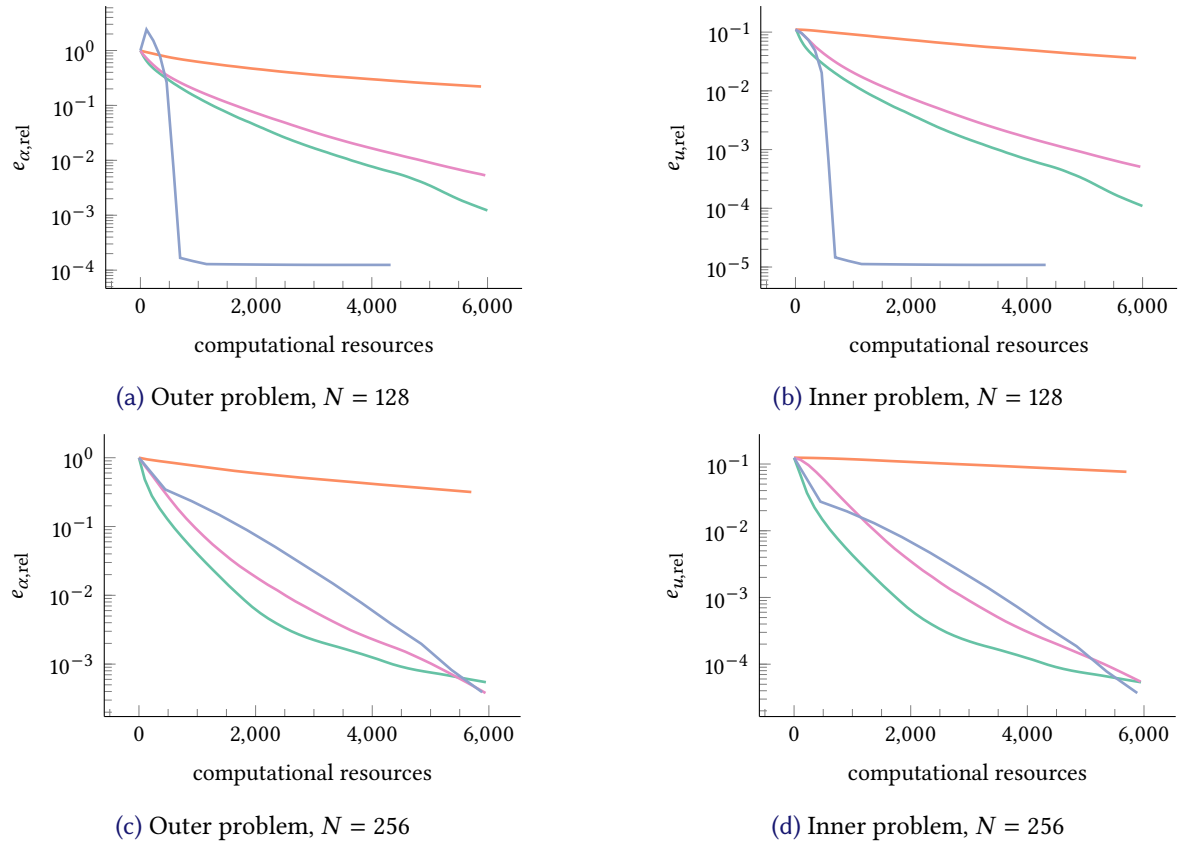
For performance comparison, we need estimates $\tilde{\alpha}$ and $\tilde{u}$ of optimal $\hat{\alpha}$ and $\hat{u} = S_u(\hat{\alpha})$. For denoising we find them by searching for the one-dimensional variable α on a regular grid and recursively subdividing until node spacing goes below 10^{-5}. As $\tilde{u}$, we take an estimate of $S_u(\tilde{\alpha})$ obtained with 25000 steps of the implicit base line method. We visualise so obtained $\tilde{\alpha}$ and $J \circ S_u$ in Figure 3. For the higher-dimensional deconvolution problem, such a scan is not feasible. Instead, we obtain the comparison estimates by running the implicit method from a very good initial iterate until computational resources (CPU time) value of 6000 for $N = 32$ and 10000 for $N = 128$. Specifically, we initialise the kernel parameters $(\alpha_2, \alpha_3, \alpha_4)$ as those used for generating the data z, and the regularization parameter $\alpha_1 = 0.045$ for $N = 32$ and $\alpha_1 = 0.02$ for $N = 128$, the latter found by trial and error. This initialisation is different from that used for the actual numerical experiments; see above. Our experiments indicate that the other methods approach $\tilde{\alpha}$ so obtained faster than the implicit method itself, providing some justification for the choice.

With these solution estimates we define the inner and outer relative errors

$$e_{\alpha,\mathrm{rel}} := \frac{\|\tilde{\alpha} - \alpha^k\|}{\|\tilde{\alpha}\|} \quad \text{and} \quad e_{u,\mathrm{rel}} := \frac{\|\tilde{u} - u^k\|}{\|\tilde{u}\|}.$$

4.5 RESULTS

We report performance in Figures 4 and 5 and the image data and reconstructions in Figures 1 and 2. Figure 5 indicates that for deconvolution the FIFB significantly outperforms the other methods. The outer variable converges much faster than for other evaluated methods despite the fact that the inner variable especially with $N = 32$ stays some distance away from $\tilde{u}$. However, as the dashed line indicates, the exact solution $S_u(\alpha^k)$ of the inner problem for the corresponding outer iterate, shows clear signs of convergence. (The few "spikes" in the graph for $N = 128$ temporarily have the regularisation parameter α_0^k much closer to zero than $\tilde{\alpha}_0$.) This observation justifies the intuition that the inner problem does



(a) Outer problem, $N = 128$

(b) Inner problem, $N = 128$

(c) Outer problem, $N = 256$

(d) Inner problem, $N = 256$

Figure 4: Denoising performance. The graphs correspond to the FIFB (——), FEFB (——), AID (——), and implicit method (——).

not need to be solved to high accuracy to obtain convergence for the outer problem; that such accurate solutions can even be detrimental to convergence. The exact solution of the adjoint equation in both the implicit method and the FEFB causes them to be too slow to make any meaningful progress. The denoising experiments of Figure 4 likewise suggest that the FIFB is *initially* the best performing algorithm, although the implicit method and the AID catch up later on the denoising problem. On the small denoising problem ($N = 128$), the implicit method is significantly faster than any oher method. Overall, and for practical purposes, nevertheless, the FIFB appears to perform the best.

APPENDIX A PROX-σ-CONTRACTIVITY

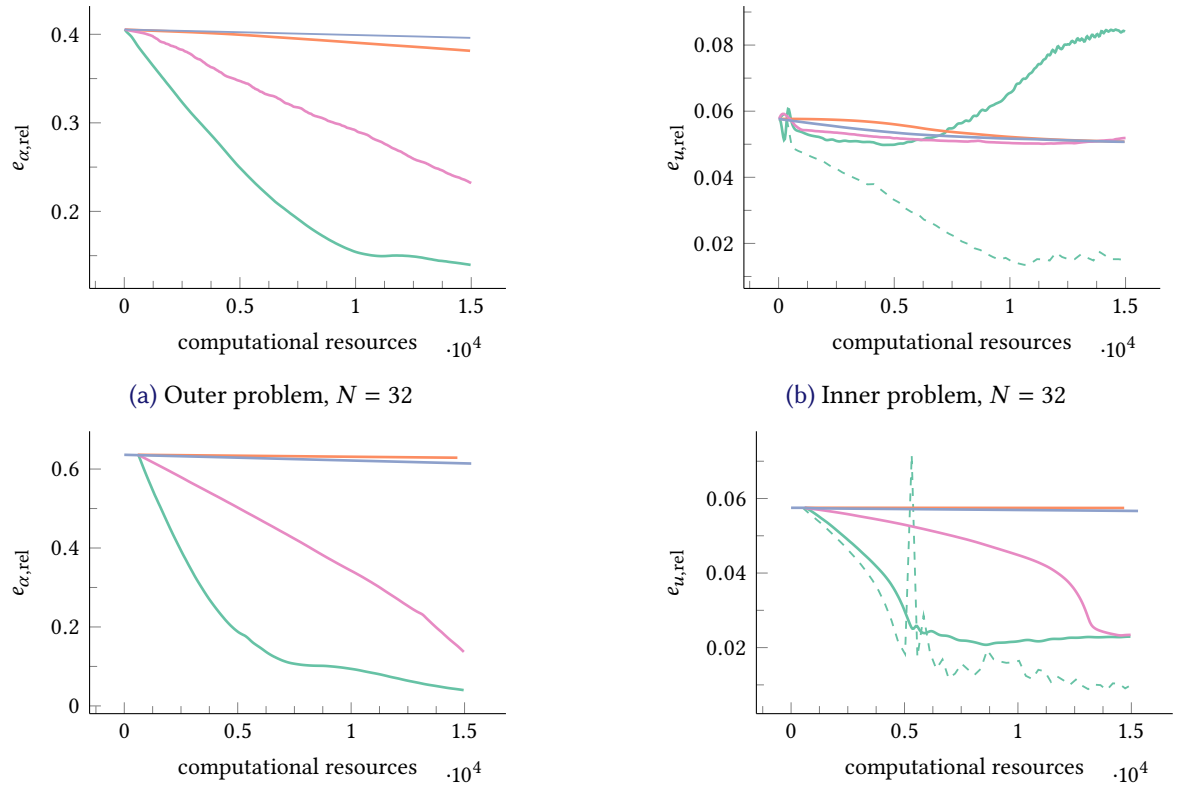
In the next three theorems we verify prox-σ-contractivity (Assumption 2.1) for some common cases.

The next theorem readily extends to $R = \beta\|\cdot\|_1 + \delta_{[0,\infty)^n}$ on $\mathbb{R}^n$ and products sets A since the proximal mapping of R is independent for each coordinate.

Theorem A.1 (prox-σ-contractivity of positivity-constrained soft-thresholding). *Let $\sigma, \beta > 0$ and $R = \beta|\cdot|_1 + \delta_{[0,\infty)}$ on $\mathbb{R}$. Then R is prox-σ-contractive at any $\widehat{\alpha} \geq \max\{0, \sigma(q+\beta)\}$ for any $q \in \mathbb{R}^n$ with any factor $0 < C_R < \sigma^{-1}$ within*

$$A = \left[\max\left\{0, \widehat{\alpha} - \frac{\widehat{\alpha} - \sigma(q+\beta)}{1 - \sigma C_R}\right\}, \infty\right).$$

In particular, if $\widehat{\alpha} \in \operatorname{dom} R = [0,\infty)$ and $-q = \beta \in \partial R(\widehat{\alpha})$, then R is locally prox-σ-contractive at $\widehat{\alpha}$ with



(a) Outer problem, $N = 32$

(b) Inner problem, $N = 32$

(c) Outer problem, $N = 128$

(d) Inner problem, $N = 128$

Figure 5: Deconvolution performance. The graphs correspond to the FIFB (——), FEFB (——), AID (——), and implicit method (——). For the inner problem, we additionally plot the relative error of $S_u(\alpha^k)$ for the FIFB (- - -).

any factor $0 < C_R < \sigma^{-1}$ *within*

$$A = [0, \infty) .$$

Proof. We have (see, e.g., [12])

$$D_{\sigma R}(\alpha) = \operatorname{prox}_{\sigma R}(\alpha - \sigma q) - \alpha = \begin{cases} -\sigma(q+\beta), & \alpha \geq \sigma(q+\beta), \\ -\alpha, & \alpha < \sigma(q+\beta). \end{cases}$$

We have by assumption $\widehat{\alpha} \geq \sigma(q+\beta)$. If also $\alpha \geq \sigma(q+\beta)$, we have $D_{\sigma R}(\alpha) - D_{\sigma R}(\widehat{\alpha}) = 0$, which satisfies the required inequality.

Suppose then that $\alpha < \sigma(q+\beta)$. Since $D_{\sigma R}(\widehat{\alpha}) = -\sigma(q+\beta)$, we need to show that

$$|D_{\sigma R}(\alpha) + \sigma(q+\beta)| \leq \sigma C_R(\widehat{\alpha} - \alpha). \tag{A.1}$$

We have $D_{\sigma R}(\alpha) = -\alpha$, so (A.1) rearranges as

$$-\alpha + \sigma(q+\beta) \leq \sigma C_R(\widehat{\alpha} - \alpha).$$

Since $1 > \sigma C_R$, this inequality can be rearranged as the condition $\alpha \geq \widehat{\alpha} - \frac{\widehat{\alpha} - \sigma(q+b)}{1-\sigma C_R}$. Any $\alpha \in A$ satisfies this bound.

Let then $-q = \beta \in \partial R(\widehat{\alpha})$. Since $\widehat{\alpha} \geq 0$, we have $\widehat{\alpha} \geq \sigma(q+\beta) = 0$. Since $\frac{\widehat{\alpha} - \sigma(q+\beta)}{1-\sigma C_R} = \frac{\widehat{\alpha}}{1-\sigma C_R} \geq \widehat{\alpha}$, the claimed simpler expression for A follows from the general. □

Similarly to the previous result, the restriction $q = 0$ in the next theorem on projections to a convex set C forbids stricly complementary cases of $\widehat{\alpha} \in \operatorname{bd} C$, i.e., we cannot have $0 \neq -q \in N_C(\widehat{\alpha}) := \partial\delta_C(\widehat{\alpha})$.

Theorem A.2 (prox-σ-contractivity of projections). *Let $\sigma > 0$ and $R = \delta_C$ for a convex and closed $C \subset \mathbb{R}^n$. Then R is prox-σ-contractive at any $\widehat{\alpha} \in C$ for $q = 0$ within $A = C$ with any factor $C_R > 0$.*

Proof. We have $\operatorname{prox}_{\sigma R} = \operatorname{proj}_C$ for the Euclidean projection onto C. Since $\alpha, \widehat{\alpha} \in C = \operatorname{dom} R$ and $q = 0$, we have $\alpha = \operatorname{proj}_C(\alpha) = \operatorname{proj}_C(\alpha - \sigma q)$, and likewise for $\widehat{\alpha}$. The claim is now immediate. □

Example A.3 (ReLu). The proximal mapping of $\delta_{[0,\infty)}$ is known as the rectifier linear unit activation function (ReLu). By the above theorem, it is prox-σ-contractive at any $\widehat{\alpha} \geq 0$ for $q = 0$.

Theorem A.4 (prox-σ-contractivity of smooth functions). *Let $\sigma, \beta > 0$ and $R : \mathbb{R}^n \to \mathbb{R}$ be convex with Lipschitz gradient. Then R is prox-σ-contractive at any $\widehat{\alpha} \in \mathbb{R}^n$ for any $q \in \mathbb{R}^n$ within $A = \mathbb{R}^n$ with the factor $C_R = L_{\nabla R}$ the Lipschitz factor of ∇R.*

Proof. Write $p(\alpha) := \operatorname{prox}_{\sigma R}(t(\alpha))$ and $t(\alpha) := \alpha - \sigma q$. According to the definition of the proximal operator, $0 = \sigma\nabla R(p(\alpha)) + p(\alpha) - t(\alpha)$. Hence $p(\alpha) - \alpha = -\sigma[q + \nabla R(p(\alpha))]$ which yields for any $\alpha \in \mathbb{R}^n$, as required,

$$\begin{aligned}
\|D_{\sigma R}(\alpha) - D_{\sigma R}(\widehat{\alpha})\| &= \|[p(\alpha) - \alpha] - [p(\widehat{\alpha}) - \widehat{\alpha}]\| \\
&= \sigma\|\nabla R(p(\alpha)) - \nabla R(p(\widehat{\alpha}))\| \\
&\leq \sigma L_{\nabla R}|||p(\alpha) - p(\widehat{\alpha})||| \\
&\leq \sigma L_{\nabla R}\|t(\alpha) - t(\widehat{\alpha})\| \\
&= \sigma L_{\nabla R}\|\alpha - \widehat{\alpha}\|.
\end{aligned}$$

□

APPENDIX B A NORM ON OPERATORS ON SEPARABLE HILBERT SPACES

We show basic properties of the inner product defined and norm induced by (2.4b).

Theorem B.1. *Let U be a Hilbert space and $\mathscr{A}$ a separable Hilbert space. On $\mathbb{L}(U;\mathscr{A})$ define $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ and $|||\cdot|||$ according to (2.4b). Then*

(i) $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is an inner product and $|||\cdot|||$ a norm on $\mathbb{L}(U;\mathscr{A})$.
(ii) For $M \in \mathbb{L}(U;U)$ and $p \in \mathbb{L}(U;\mathscr{A})$, we have $|||pM||| \leq |||p|||\|M\|_{\mathbb{L}(U;U)}$.
(iii) The operator norm on $\mathbb{L}(U;\mathscr{A})$ satisfies $\|\cdot\|_{\mathbb{L}(U;\mathscr{A})} \leq |||\cdot|||$.

Proof. (i) Clearly $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is bilinear and symmetric. Also $|||p||| = \langle\!\langle p, p \rangle\!\rangle \geq 0$ for all $p \in P$. To prove that $|||p||| > 0$ for $p \neq 0$, we observe that the contrary implies $\|p^*\varphi_i\| = 0$ for all $i \in I$. Since $\{\varphi_i\}_{i\in I}$ is a basis for $\mathscr{A}$, this implies $p^* = 0$, hence $p = 0$.

(ii) We have $|||pM|||^2 = \sum_{i\in I}\|M^*p^*\varphi_i\|^2 \leq \sum_{i\in I}\|M^*\|^2_{\mathbb{L}(U;U)}\|p^*\varphi_i\|^2 = \|M\|^2_{\mathbb{L}(U;U)}|||p|||^2$.

(iii) Let $p \in \mathbb{L}(U;\mathscr{A})$. Then $\|p\|_{\mathbb{L}(U;\mathscr{A})} = \|p^*\|_{\mathbb{L}(\mathscr{A};U)} = \sup_{\alpha\in\mathscr{A},\|\alpha\|=1}\|p^*\alpha\|_U$. Since $\{\varphi_i\}_{i\in I}$ is an orthonormal basis for $\mathscr{A}$, we can write $\alpha = \sum_{i\in I} a_i\varphi_i$ for some $a_i \in \mathbb{R}$ with $\sum_{i\in I} a_i^2 = 1$. Thus $\|p^*\alpha\|_U^2 = \sum_{i\in I}\sum_{\alpha\in I} a_i a_j\langle p^*\varphi_i, p^*\varphi_j\rangle_U \leq \sum_{i\in I}\left(\sum_{\alpha\in I} a_j^2\right)\|p^*\varphi_i\|_U^2 = |||p|||^2$, where the last inequality uses Young's inequality. The claim follows. □

REFERENCES

[1] G. B. Allende and G. Still, Solving bilevel programs with the KKT-approach, *Mathematical Programming* 138 (2012), 309–332, doi:10.1007/s10107-012-0535-x.

[2] D. Aussel and C. S. Lalitha (eds.), *Generalized Nash Equilibrium Problems, Bilevel Programming and MPEC*, Springer Singapore, 2017, doi:10.1007/978-981-10-4774-9.

[3] K. Bai and J. J. Ye, Directional Necessary Optimality Conditions for Bilevel Programs, *Mathematics of Operations Research* (2021), doi:10.1287/moor.2021.1164. To appear (published online).

[4] J. F. Bard and J. E. Falk, An explicit solution to the multi-level programming problem, *Computers & Operations Research* 9 (1982), 77–100, doi:10.1016/0305-0548(82)90007-7.

[5] 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.

[6] L. Calatroni, J. C. De Los Reyes, and C. B. Schönlieb, Dynamic Sampling Schemes for Optimal Noise Learning Under Multiple Nonsmooth Constraints, in *Proceedings of the 26th IFIP TC 7 Conference on System Modeling and Optimization, Klagenfurt, Austria*, 2014.

[7] 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.

[8] A. Chambolle and T. Pock, Learning Consistent Discretizations of the Total Variation, *SIAM Journal on Imaging Sciences* 14 (2021), 778–813, doi:10.1137/20m1377199.

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

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

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

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

[13] 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.

[14] 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.

[15] J. C. De Los Reyes, C. B. Schönlieb, and T. Valkonen, The structure of optimal parameters for image restoration problems, *Journal of Mathematical Analysis and Applications* 434 (2016), 464–500, doi:10.1016/j.jmaa.2015.09.023, arXiv:1505.01953.

[16] 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.

[17] 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.

[18] J. C. De Los Reyes and D. Villacís, Bilevel Imaging Learning Problems as Mathematical Programs with Complementarity Constraints, 2021, arXiv:2110.02273.

[19] S. Dempe, *Foundations of Bilevel Programming*, Nonconvex Optimization and Its Applications, Springer US, 2006.

[20] S. Dempe, Bilevel optimization: theory, algorithms and applications, preprint 2018-11, TU Bergakademie Freiberg, Fakultät für Mathematik und Informatik, 2018, http://www.optimization-online.org/DB_FILE/2018/08/6773.pdf.

[21] S. Dempe, V. Kalashnikov, G. A. Pérez-Valdés, and N. Kalashnykova, *Bilevel Programming Problems*, Energy Systems, Springer Berlin Heidelberg, 2015, doi:10.1007/978-3-662-45827-3.

[22] S. Dempe and A. B. Zemkoho, The bilevel programming problem: reformulations, constraint qualifications and optimality conditions, *Mathematical Programming* 138 (2012), 447–473, doi:10.1007/s10107-011-0508-5.

[23] M. J. Ehrhardt and L. Roberts, Inexact Derivative-Free Optimization for Bilevel Learning, *Journal of Mathematical Imaging and Vision* (2021), doi:10.1007/s10851-021-01020-8.

[24] J. E. Falk and J. Liu, On bilevel programming, Part I: General nonlinear cases, *Mathematical Programming* 70 (1995), 47–72, doi:10.1007/bf01585928.

[25] J. Fliege, A. Tin, and A. Zemkoho, Gauss–Newton-type methods for bilevel optimization, *Computational Optimization and Applications* 78 (2021), 793–824, doi:10.1007/s10589-020-00254-3.

[26] 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/.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] Y. Jiang, X. Li, C. Huang, and X. Wu, Application of particle swarm optimization based on CHKS smoothing function for solving nonlinear bilevel programming problem, *Applied Mathematics and Computation* 219 (2013), 4332–4339, doi:10.1016/j.amc.2012.10.010.

[37] 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.

[38] 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.

[39] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang, A Generic First-Order Algorithmic Framework for Bi-Level Programming Beyond Lower-Level Singleton, in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh (eds.), volume 119 of Proceedings of Machine Learning Research, PMLR, 2020, 6305–6315.

[40] Z. Luo, J. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, 1996.

[41] P. Mehlitz and A. B. Zemkoho, Sufficient Optimality Conditions in Bilevel Programming, *Mathematics of Operations Research* (2021), doi:10.1287/moor.2021.1122.

[42] P. Ochs, R. Ranftl, T. Brox, and T. Pock, Techniques for Gradient-Based Bilevel Optimization with Non-smooth Lower Level Problems, *Journal of Mathematical Imaging and Vision* 56 (2016), 175–194, doi:10.1007/s10851-016-0663-7.

[43] S. Sabach and S. Shtern, A First Order Method for Solving Convex Bilevel Optimization Problems, *SIAM Journal on Optimization* 27 (2017), 640–660, doi:10.1137/16m105592x.

[44] Y. Shehu, P. T. Vuong, and A. Zemkoho, An inertial extrapolation method for convex simple bilevel optimization, *Optimization Methods and Software* 36 (2019), 1–19, doi:10.1080/10556788.2019.1619729.

[45] 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.

[46] A. Z. Stephan Dempe (ed.), *Bilevel Optimization: Advances and Next Challenges*, Springer International Publishing, 2020.

[47] E. Suonperä, Codes for "Linearly convergent bilevel optimization with single-step inner methods", 2023, doi:10.5281/zenodo.7974062.

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

[49] 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.

[50] T. Valkonen, First-order primal-dual methods for nonsmooth nonconvex optimisation, in *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, K. Chen, C. B. Schönlieb, X. C. Tai, and L. Younes (eds.), Springer, Cham, 2021, doi:10.1007/978-3-030-03009-4_93-1, arXiv:1910.00115.

[51] 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.

[52] A. Wächter, *An Interior Point Algorithm for Large-Scale Nonlinear Optimization with Applications in Process Engineering*, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.

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

[54] J. J. Ye and D. L. Zhu, Optimality conditions for bilevel programming problems, *Optimization* 33 (1995), 9–27, doi:10.1080/02331939508844060.

[55] A. B. Zemkoho, Solving Ill-posed Bilevel Programs, *Set-valued and Variational Analysis* 24 (2016), 423–448, doi:10.1007/s11228-016-0371-x.