DIFFERENTIAL ESTIMATES FOR FAST FIRST-ORDER MULTILEVEL NONCONVEX OPTIMISATION

Neil Dizon* Tuomo Valkonen†

Abstract With a view on bilevel and PDE-constrained optimisation, we develop iterative estimates $\widetilde{F'}(x^k)$ of $F'(x^k)$ for compositions $F := J \circ S$, where S is the solution mapping of the inner optimisation problem or PDE. The idea is to form a single-loop method by interweaving updates of the iterate x^k by an outer optimisation method, with updates of the estimate by single steps of standard optimisation methods and linear system solvers. When the inner methods satisfy simple tracking inequalities, the differential estimates can almost directly be employed in standard convergence proofs for general forward-backward type methods. We adapt those proofs to a general inexact setting in normed spaces, that, besides our differential estimates, also covers mismatched adjoints and unreachable optimality conditions in measure spaces. As a side product of these efforts, we provide improved convergence results for nonconvex Primal-Dual Proximal Splitting (PDPS).

1 INTRODUCTION

First-order methods are slow. To be precise, they require a high number of iterations, but if those iterations are fast, they have the chance to practically overpower second-order methods with expensive iterations. In PDE-constrained or bilevel optimisation, the steps of basic first-order methods are very expensive, involving the solution of the inner problem or PDE and its adjoint. To make first-order methods fast, it is, therefore, imperative to reduce the cost of solving these subproblems—or to not solve them at all.

Consequently, especially in the machine learning community, an interest has surfaced in *single-loop* methods for bilevel optimisation; see [24] and references therein. Many of these methods are very specific constructions. In [14] we started work on a more general approach to PDE-constrained optimisation: we showed that on each step of an outer primal-dual optimisation method, we can take *single steps* of standard linear system splitting schemes for the PDE constraint and its adjoint, and still obtain a convergent method that is computationally significantly faster than solving the PDEs exactly. In [25] we then presented an approach to bilevel optimisation that allowed general inner and adjoint algorithms that satisfy certain *tracking inequalities*. These were proved for standard splitting schemes for the adjoint equation, and for forward-backward splitting and the Primal-Dual Proximal Splitting (PDPS) of [3] for the inner problem. The overall analysis was still tied to bilevel optimisation in Hilbert spaces, with forward-backward splitting as the outer optimisation method.

Now, in Section 2, we show in *general normed spaces* that we can approximate in a single-loop fashion the differentials of compositions $F = J \circ S$, given abstract inner and adjoint algorithms for

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

*School of Mathematics and Statistics, University of New South Wales, Sydney, Australia. n.dizon@unsw.edu.au, ORCID: 0000-0001-8664-2255

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

S. In contrast to [25] and, indeed, all single-loop bilevel optimisation methods that we are aware of, our approach can also work with the adjoint dimension reduction trick typically employed in PDE-constrained optimisation. We show that, subject to additive error terms with a bounded sum, the differential estimates $\widetilde{F'}(x^k)$ satisfy standard smoothness properties, such as Lipschitz differential and the two- and three-point descent inequalities [27, 6].

We are interested in applying the differential estimates $\widetilde{F'}(x^k)$ in a normed space X to the solution of composite optimisation problems with optimality conditions

$$0 \in F'(x) + \partial G(x) + \Xi x \tag{1.1}$$

for G convex but possibly nonsmooth, and F typically nonconvex but smooth. The operator $\Xi \in \mathbb{L}(X; X^*)$ is skew-adjoint, allowing the modelling of primal-dual problems, and treating the PDPS and Douglas–Rachford splitting as generalised forward-backward splitting methods [6]. To facilitate the analysis of such methods, we introduce in Section 3 operator-relative variants of the descent inequality. Based on this, in Section 4, we then prove various forms of convergence.

Through our approach to inexactness, besides gradient estimates for multilevel problems, we can model mismatched adjoints [15], and difficult-to-solve-exactly optimality conditions in measure spaces [30]. We also adopt the approach of [30] to optimisation in normed spaces: instead of Bregman divergences, we construct an inner product structure with a self-adjoint $M \in \mathbb{L}(X; X^*)$. Our work is related to the study of gradient oracles for smooth convex optimisation in [8], and for nonconvex composite optimisation in [9, 18], both in finite-dimensional Euclidean spaces. Based on sufficent descent and the Kurdyka–Łojasiewicz property, [19] also study inexact methods in $\mathbb{R}^n$. Moreover, [2] introduce approaches to control model inexactness in proximal trust region methods, and [23] in non-single-loop gradient methods for bilevel optimisation.

Not content to merely adapt existing proofs to inexact steps and normed spaces, we also present some improvements, especially for the nonconvex PDPS of [26]. We do, however, treat a slightly simplified problem. The original PDPS of [3] applies to $\min_x g(x) + h(K(x))$ with K linear and g and h convex. The extension of [26] allows K to be nonlinear. It is further analysed in [4, 5, 16, 10], with a simplified review of the different variants in Banach space in [28]. An alternative extension in [17] allows g and h to be semiconvex. We, instead, consider the special case $\min_x f(x) + g(x) + h(Kx)$ with K linear, but f nonconvex, and show that the values of the convex envelope of the objective function at ergodic iterates locally converge to a minimum.

We do not include numerical results, as that has already been done in [14, 25]. Through our work, the specific algorithms presented therein can be understood through a clean and generic differential estimation approach.

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. X^* stands for the dual space. When X is Hilbert, we identify X^* with X. We write $\langle x, y \rangle$ for an inner product, $\langle x^* | x \rangle_{X^*,X}$ for a dual product. We call $M \in \mathbb{L}(X; X^*)$ *self-adjoint* if the restriction $M^*|X = M$, and *positive semi-definite* if $\langle x | Mx \rangle_{X,X^*} \geq 0$ for all $x \in X$. If both hold, we set $\|x\|_M := \sqrt{\langle Mx | x \rangle}$, and write $\mathbb{O}_M(x, r)$ for the radius-r open ball at x in the M-(semi-)norm. We also write $M \geq N$ if $M - N$ is positive semi-definite. We extensively use the vectorial Young's inequality

$$\langle x^* | x \rangle_{X^*,X} \leq \frac{a}{2}\|x\|_X^2 + \frac{1}{2a}\|x^*\|_{X^*}^2 \qquad \text{for all } x \in X,\ x^* \in X^*\ a > 0.$$

For $F : X \to \mathbb{R}$, we write $DF(x)$ for the Gâteaux and $F'(x) \in X^*$ for the Fréchet derivative at x, if they exist. If X is Hilbert, $\nabla F(x) \in X$ stands for the Riesz representation of $F'(x)$, i.e., the gradient. For partial derivatives, we use the notation $F^{(x)}(u, x)$. We also write $\operatorname{sub}_c F := \{x \in X \mid F(x) \leq c\}$ for the

c-sublevel set. With $\overline{\mathbb{R}} := [-\infty, \infty]$, for a convex $G : X \to \overline{\mathbb{R}}$, we write $\operatorname{dom} G$ for the effective domain, $\partial G(x)$ for the subdifferential at x, and $G^* : X^* \to \overline{\mathbb{R}}$ for the Fenchel conjugate. When X is a Hilbert space, we write prox_F for the proximal map and, with a slight abuse of notation, identify $\partial G(x)$ with the set of Riesz representations of its elements.

2 TRACKING ESTIMATE RECURSION

Let $J : U \to \mathbb{R}$ and $S_u : X \to U$ be Fréchet differentiable on normed spaces X and U. We consider the functional

$$F(x) = J(S_u(x)).$$

We estimate $S_u(x^k)$ by $u^{k+1} \in U$, and $S_u'(x^k)$ by $p^{k+1} \in \mathbb{L}(X; U)$, that is, we estimate

$$F'(x^k) = J'(S_u(x^k))S_u'(x^k) \quad \text{by} \quad \widetilde{F'}(x^k) = J'(u^{k+1})p^{k+1}.$$

When X is Hilbert, we write $\widetilde{\nabla F}(x^k)$ for the Riesz representation of $\widetilde{F'}(x^k)$. Our goal is to derive, in Section 2.3, smoothness estimates for this approximation, given *tracking estimates* for u^{k+1} and p^{k+1}, derived from the contractivity of splitting methods as in [14, 25]. We state and provide examples of those tracking estimates in Section 2.1. These are followed by several technical lemmas in Section 2.2.

Although $\widetilde{F'}(x^k)$ will have the above structure, we want to avoid constructing $p^{k+1} \approx S_u'(x^k) \in \mathbb{L}(X; U)$ directly due to its high dimensionality. Instead, we seek to only construct the necessary projections through a lower-dimensional variable w^{k+1}.

Example 2.1 (Adjoint equations). Suppose $S_u(x)$ arises from the satisfaction of

$$0 = T(S_u(x), x) \quad \text{for a} \quad T : U \times X \to W_* \quad \text{with} \quad W_* \text{ a normed space,} \tag{2.1}$$

modelling, e.g., a PDE or the first-order optimality conditions of an inner optimisation problem, both parametrised by x. By implicit differentiation, subject to sufficient differentiability and (2.1) holding in a neighbourhood of x, we obtain the *basic adjoint*

$$T^{(u)}(S_u(x), x)S_u'(x) + T^{(x)}(S_u(x), x) = 0 \in W := (W_*)^*, \tag{2.2}$$

where $S_u'(x) \in \mathbb{L}(X; U)$, $T^{(u)}(S_u(x), x) \in \mathbb{L}(U; W)$, and $T^{(x)}(S_u(x), x) \in \mathbb{L}(X; W)$. Hence, following the derivation of adjoint PDEs in, e.g., [13, §1.6.2] or [7, §1.2], assuming $T^{(u)}(S_u(x), x)$ to be invertible, we solve from (2.2) that

$$[J \circ S_u]'(x) = J'(S_u(x))S_u'(x) = w_x T^{(x)}(S_u(x), x),$$

for a $w_x = S_w(x) \in W$ satisfying the *reduced adjoint*

$$w_x T^{(u)}(S_u(x), x) + J'(S_u(x)) = 0. \tag{2.3}$$

For $x = x^k$, we will in practise take w^{k+1} as an operator splitting approximation to

$$w^{k+1} T^{(u)}(u^{k+1}, x^k) + J'(u^{k+1}) = 0, \tag{2.4}$$

and then set

$$\widetilde{F'}(x^k) := w^{k+1} T^{(x)}(u^{k+1}, x^k) \approx J'(S_u(x^k))S_u'(x^k).$$

2.1 BASIC CONSTRUCTIONS AND ASSUMPTIONS

In the next principal assumption, typically the distances b_X, d_{X^*}, d_U, and d_W would be given by norms, X^* be the dual space of X, and the "target" $F'(x^k)$ be the Fréchet derivative of $J \circ S_u$ at x^k, but this need not be the case; they are entirely formal expressions, and the spaces are treated as sets without structure. We write squared distances as $b_X^2(x, y) := b_X(x, y)^2$.

Assumption 2.2. For spaces X, X^*, U, and W, a subset $\Omega \subset X$, an *inner solution map* $S_u : X \to U$ and an *adjoint solution map* $S_w : X \to W$, the following hold:

(i) We are given an *inner algorithm* that, on each iteration $k \geq 1$, given $\{x^n\}_{n=0}^k \subset \Omega$ and $u^k \in U$, produces $u^{k+1} \in U$ satisfying

$$\kappa_u d_U(u^{k+1}, S_u(x^k)) \leq d_U(u^k, S_u(x^{k-1})) + \pi_u b_X(x^k, x^{k-1})$$

for some $\pi_u > 0$, $\kappa_u > 1$, and distance expressions d_U and b_X.

(ii) We are given an *adjoint algorithm* that, on each iteration $k \geq 1$, given $\{(x^n, u^{n+1})\}_{n=0}^k \subset \Omega \times U$, and $w^k \in W$, produces $w^{k+1} \in W$ satisfying

$$\kappa_w d_W(w^{k+1}, S_w(x^k)) \leq d_W(w^k, S_w(x^{k-1})) + \mu_u d_U(u^{k+1}, S_u(x^k)) + \pi_w b_X(x^k, x^{k-1})$$

for some $\mu_u, \pi_w > 0$, $\kappa_w > 1$, and a distance expression d_W.

(iii) We are given a *differential transformation* that, on each iteration $k \in \mathbb{N}$, given $u^{k+1} \in U$ and $w^{k+1} \in W$, produces $\widetilde{F'}(x^k) \in X^*$ that, for a *target* $F'(x^k) \in X^*$, satisfies for some $\alpha_u, \alpha_w \geq 0$ and a distance expression d_{X^*} the bound

$$d_{X^*}(\widetilde{F'}(x^k), F'(x^k)) \leq \alpha_u d_U(u^{k+1}, S_u(x^k)) + \alpha_w d_W(w^{k+1}, S_w(x^k)).$$

The inner and adjoint tracking conditions (i) and (ii) are parameter change aware contractivity conditions for the inner and adjoint algorithms: if $x^k = x^{k-1}$, the former is simply a contractivity condition. The condition (iii) allow converting the construction error of $\widetilde{F'}(x^k)$ to the tracking errors of the inner and adjoint algorithms.

We next provide brief examples of methods that satisfy the parts of Assumption 2.2; for (i) and (ii) the proofs are in [25]. In all cases S_u is determined by (2.1) for a prescribed T. In the first examples of inner problems and algorithms, X is a normed space and U is a Hilbert space.

Example 2.3 (Inner algorithm: forward-backward). Let $T(u, x) = \nabla f(u; x) + \nabla g(u; x)$ for f and g convex in u, and differentiable in (u, x); $\nabla f(\,\cdot\,; x)$ L-Lipschitz, and $g(\,\cdot\,; x)$ γ-strongly convex, both uniformly in x. If $S_u(x) = \arg\min_u [f + g](u; x)$ is Lipschitz in Ω, then the forward-backward splitting updates $u^{k+1} := \mathrm{prox}_{\tau g(\,\cdot\,; x^k)}(u^k - \tau \nabla f(u^k; x^k))$ satisfy Assumption 2.2 (i) when $\tau L \leq 2$ [25, Theorem 3.4].

Example 2.4 (Inner algorithm: primal-dual). Represent the Fenchel–Rockafellar primal-dual optimality conditions of $\min_z f(z; x) + g^*(Kz; x)$, as roots u of

$$T(u, x) = (\nabla f(z; x) + K^* y, \nabla g(y; x) - Kz) \quad \text{where} \quad u = (z, y) \in U = Z \times Y$$

for $K \in \mathbb{L}(Z; Y^*)$ linear and bounded, both f and g convex in the first parameter, differentiable in both parameters; and $g(\,\cdot\,; x)$ γ-strongly convex uniformly in x. If $S_u(x) = T^{-1}(\,\cdot\,; x)(0)$ is Lipschitz in Ω, then the PDPS updates [3]

$$z^{k+1} = \mathrm{prox}_{\tau f(\,\cdot\,; x^k)}(z^k - \tau K^* y^k) \quad \text{and} \quad y^{k+1} = \mathrm{prox}_{\sigma g(\,\cdot\,; x^k)}(y^k + \sigma K(2z^{k+1} - z^k))$$

satisfy Assumption 2.2 (i) when $\tau\sigma\|K\| \leq 1$ [25, Theorem 3.6].

The next example covers PDE-constrained optimisation.

Example 2.5 (Inner algorithm: linear system splitting). Let $T(u, x) = A_x u - b_x$ with both $A_x \in \mathbb{L}(U; U)$ and $b_x \in U$ Lipschitz in x. If $S_u(x) = A_x^{-1} b_x$ is Lipschitz in Ω, then splitting $A_x = N_x + M_x$ per the Jacobi or Gauss–Seidel schemes, the updates $u^{k+1} = N_{x^k}^{-1}(b_{x^k} - M_{x^k} u^k)$ satisfy Assumption 2.2 (i) subject to uniform versions of standard convergence conditions for these schemes [25, Examples 4.3 and 4.4].

The following two examples treat adjoint methods for Examples 2.3 to 2.5.

Example 2.6 (Basic adjoint: linear system splitting). Let S_u be given by (2.1), and $S_w = S_p = S_u'$ through the basic adjoint (2.2). If $T|U \times \Omega$ and $S_u|\Omega$ are Lipschitz-continuously differentiable, then, subject to uniformised standard convergence conditions, Jacobi or Gauss–Seidel updates applied to $0 = T^{(u)}(u^{k+1}, x^k)p^{k+1} + T^{(x)}(u^{k+1}, x^k)$, satisfy Assumption 2.2 (ii) with $w^k = p^k$ [25, Theorem 4.9].

Example 2.7 (Reduced adjoint: linear system splitting). With S_u given by (2.1), define S_w by (2.3). If $T|U \times \Omega$, $S_u|\Omega$, and J are Lipschitz-continuously differentiable, then Jacobi or Gauss–Seidel updates applied to (2.4) satisfy Assumption 2.2 (ii) subject to uniformised standard convergence conditions. Although not proved in [25], this claim follows similarly to the basic adjoint, as both are linear systems of similar form, where J' in the reduced adjoint takes the place of $T^{(x)}$ in the basic adjoint.

The final two examples treat the construction of the differential estimate.

Example 2.8 (Differential transformation: basic adjoint). As in Example 2.6, take $W = \mathbb{L}(X; U)$ and let $S_w(x) = S_u'(x)$ be determined by the *basic adjoint* (2.2). Suppose

$$N_{J'} := \sup\{\|J'(S_u(x))\|_{U^*} \mid x \in \Omega\} < \infty, \quad N_{S_u'} := \sup\{\|S_u'(x)\|_{\mathbb{L}(X;U)} \mid x \in \Omega\} < \infty,$$

and that J' is $L_{J'}$-Lipschitz. Let $\widetilde{F'}(x^k) = J'(u^{k+1})p^{k+1}$ for $p^{k+1} \approx S_u'(x^k)$ generated, for example, by the methods of Example 2.6. Then Assumption 2.2 (iii) holds due to

$$\begin{aligned}
\|\widetilde{F'}(x^k) - F'(x^k)\|_{X^*} &= \|J'(u^{k+1})p^{k+1} - J'(S_u(x^k))S_u'(x^k)\|_{X^*} \\
&\leq \|J'(u^{k+1})[p^{k+1} - S_u'(x^k)]\|_{X^*} + \|[J'(u^{k+1}) - J'(S_u(x^k))]S_u'(x^k)\|_{X^*} \\
&\leq \|p^{k+1} - S_u'(x^k)\|_{\mathbb{L}(X;U)}\|J'(u^{k+1})\|_{U^*} + L_{\nabla J}\|S_u'(x^k)\|_{\mathbb{L}(X;U)}\|u^{k+1} - S_u(x^k)\|_U \\
&\leq N_{\nabla J}\|p^{k+1} - S_u'(x^k)\|_{\mathbb{L}(X;U)} + L_{\nabla J} N_{\nabla S_u}\|u^{k+1} - S_u(x^k)\|_U.
\end{aligned}$$

Example 2.9 (Differential transformation: reduced adjoint). Continuing from Example 2.1, take $S_w(x) = w_x \in W$ as a solution of the (2.3), that is $w_x T^{(u)}(S_u(x), x) + J'(S_u(x)) = 0$. Suppose $T^{(x)}(\,\cdot\,, x)$ is $L_{T^{(x)};u}$-Lipshitz for all $x \in \Omega$ with both

$$M_{T^{(x)}} := \sup\{\|T^{(x)}(u, x)\| \mid u \in U, x \in \Omega\} < \infty \quad \text{and} \quad N_{S_w} := \sup\{\|S_w(x)\| \mid x \in \Omega\} < \infty.$$

Take $\widetilde{F'}(x^k) := w^{k+1}T^{(x)}(u^{k+1}, x^k)$ for w^{k+1} produced, for example, by the methods of Example 2.7.

Then the differential transformation Assumption 2.2 (iii) holds due to

$$
\begin{aligned}
\|\widetilde{F}'(x^k) - F'(x^k)\|_{X^*} &= \|w^{k+1}T^{(x)}(u^{k+1}, x^k) - w_{x^k}T^{(x)}(S_u(x^k), x^k)\|_{X^*} \\
&= \|[w^{k+1} - w_{x^k}]T^{(x)}(u^{k+1}, x^k) - w_{x^k}[T^{(x)}(S_u(x^k), x^k) - T^{(x)}(u^{k+1}, x^k)]\|_{X^*} \\
&\le \|T^{(x)}(u^{k+1}, x^k)\|_{\mathbb{L}(X;W^*)}\|w^{k+1} - w_{x^k}\|_W \\
&\quad + \|w_{x^k}\|_W\|T^{(x)}(S_u(x^k), x^k) - T^{(x)}(u^{k+1}, x^k)\|_{\mathbb{L}(X;W^*)} \\
&\le N_{S_w}L_{T^{(x)};u}\|u^{k+1} - S_u(x^k)\|_U + M_{T^{(x)}}\|w^{k+1} - S_w(x^k)\|_W.
\end{aligned}
$$

2.2 TECHNICAL LEMMAS

We start with the following result on sequences of real numbers. We will later parametrise it according to the inner and adjoint tracking inequalities.

Lemma 2.10. *For some $\kappa_u, \kappa_w > 1$ and $\mu_u, \pi_u, \pi_w > 0$ suppose $b_k, c_k, d_k \ge 0$ for all $k \ge 0$ satisfy*

$$\kappa_u b_{k+1} \le b_k + \pi_u d_k \quad \text{and} \quad \kappa_w c_{k+1} \le c_k + \mu_u b_{k+1} + \pi_w d_k.$$

Then, letting $\iota_k := \sum_{m=1}^{k} \kappa_u^{-m}\kappa_w^{-(k+1-m)}$ (understanding that $\iota_0 = 0$), for all $k \ge 0$ and $\alpha_u, \alpha_w \ge 0$, we have

$$
\begin{aligned}
R^{k+1}(\alpha_u, \alpha_w) := \alpha_u b_{k+1} + \alpha_w c_{k+1} &\le (\alpha_u\kappa_u^{-k} + \alpha_w\iota_k\mu_u)b_1 + \alpha_w\kappa_w^{-k}c_1 \\
&\quad + \sum_{j=0}^{k-1}\bigl(\alpha_u\kappa_u^{-(k-j)}\pi_u + \alpha_w[\iota_{k-j}\mu_u\pi_u + \kappa_w^{-(k-j)}\pi_w]\bigr)d_{j+1}.
\end{aligned}
\tag{2.5}
$$

Proof. For $k = 1$, $b_2 \le \kappa_u^{-1}b_1 + \kappa_u^{-1}\pi_u d_1$ and $c_2 \le \kappa_w^{-1}c_1 + \kappa_w^{-1}\mu_u b_2 + \kappa_w^{-1}\pi_w d_1$ by assumption. Multiplying the former by $\alpha_u + \alpha_w\kappa_w^{-1}\mu_u$ and the latter by α_w, then summing up, observing to cancel the two instances of $\alpha_w\kappa_w^{-1}\mu_u b_2$, establishes (2.5).

We then take $k = n+1$, and proceed by induction, assuming (2.5) to hold for $k = n$. Again, $b_{n+2} \le \kappa_u^{-1}b_{n+1} + \kappa_u^{-1}\pi_u d_{n+1}$ and $c_{n+2} \le \kappa_w^{-1}c_{n+1} + \kappa_w^{-1}\mu_u b_{n+1} + \kappa_w^{-1}\pi_w d_{n+1}$ by assumption. As in the case $k = 1$, multiplying the former by $\alpha_u + \alpha_w\kappa_w^{-1}\mu_u$ and the latter by α_w, and then summing up, yields

$$
\begin{aligned}
R^{n+2}(\alpha_u, \alpha_w) = \alpha_u b_{n+2} + \alpha_w\kappa_w^{-1}c_{n+2} &\le (\alpha_u\kappa_u^{-1} + \alpha_w\kappa_w^{-1}\kappa_u^{-1}\mu_u)b_{n+1} + \alpha_w\kappa_w^{-1}c_{n+1} \\
&\quad + (\alpha_u\kappa_u^{-1}\pi_u + \alpha_w[\kappa_w^{-1}\kappa_u^{-1}\pi_u\mu_u + \kappa_w^{-1}\pi_w])d_{n+1}.
\end{aligned}
$$

The first two terms on the right-hand side equal $R^{n+1}(\alpha_u\kappa_u^{-1} + \alpha_w\kappa_w^{-1}\kappa_u^{-1}\mu_u, \alpha_w\kappa_w^{-1})$, so using (2.5) for $k = n$, we continue

$$
\begin{aligned}
R^{n+2}(\alpha_u, \alpha_w) &\le ((\alpha_u\kappa_u^{-1} + \alpha_w\kappa_w^{-1}\kappa_u^{-1}\mu_u)\kappa_u^{-n} + \alpha_w\kappa_w^{-1}\iota_n\mu_u)b_1 + \alpha_w\kappa_w^{-1}\kappa_w^{-n}c_1 \\
&\quad + \sum_{j=0}^{n-1}\bigl((\alpha_u\kappa_u^{-1} + \alpha_w\kappa_w^{-1}\kappa_u^{-1}\mu_u)\kappa_u^{-(n-j)}\pi_u + \alpha_w\kappa_w^{-1}[\iota_{n-j}\mu_u\pi_u + \kappa_w^{-(n-j)}\pi_w]\bigr)d_{j+1} \\
&\quad + (\alpha_u\kappa_u^{-1}\pi_u + \alpha_w[\kappa_w^{-1}\kappa_u^{-1}\pi_u\mu_u + \kappa_w^{-1}\pi_w])d_{n+1} \\
&= (\alpha_u\kappa_u^{-(n+1)} + \alpha_w\mu_u(\kappa_w^{-1}\kappa_u^{-(n+1)} + \kappa_w^{-1}\iota_n))b_1 + \alpha_w\kappa_w^{-(n+1)}c_1 \\
&\quad + \sum_{j=0}^{n}\bigl(\alpha_u\kappa_u^{-(n+1-j)}\pi_u + \alpha_w[(\kappa_w^{-1}\kappa_u^{-(n+1-j)} + \kappa_w^{-1}\iota_{n-j})\mu_u\pi_u + \kappa_w^{-(n+1-j)}\pi_w]\bigr)d_{j+1}.
\end{aligned}
$$

Here $\kappa_w^{-1}\kappa_u^{-(n+1-j)} + \kappa_w^{-1}\iota_{n-j} = \iota_{n+1-j}$, as by the definition of ι_{n+1}, for any $n \ge 0$,

$$\iota_{n+1} = \sum_{m=1}^{n+1}\kappa_u^{-m}\kappa_w^{-(n+2-m)} = \kappa_w^{-1}\kappa_u^{-(n+1)} + \sum_{m=1}^{n}\kappa_u^{-m}\kappa_w^{-(n+2-m)} = \kappa_w^{-1}\kappa_u^{-(n+1)} + \kappa_w^{-1}\iota_n, \tag{2.6}$$

Thus we obtain (2.5) for $k = n+1$. □

The next two, still very technical, lemmas form our core estimates. To simplify the estimates, recalling that $\kappa_u, \kappa_w > 1$, we observe that

$$p^k \iota_k \le p^{-1} k(\kappa/p)^{-(k+1)} \quad \text{for} \quad \kappa := \min(\kappa_u, \kappa_w) > 1 \text{ and any } p \in (0, \kappa). \tag{2.7}$$

Thus, by sum formulae for arithmetic-geometric progressions [11, formula 0.113],

$$\sum_{k=0}^{n-1} p^k \iota_k \le \sum_{k=0}^{\infty} p^k \iota_k \le p^{-1}(\kappa/p - 1)^{-2} = p(\kappa - p)^{-2} \quad \text{for all } n \in \mathbb{N}. \tag{2.8}$$

Lemma 2.11. *Suppose Assumption 2.2 holds and that $\{x^n\}_{n=0}^k \subset \Omega$ for a $k \in \mathbb{N}$. Then for any $x \in X$, $p \in (0, \kappa)$, and $s \in \mathbb{R}$,*

$$4s d_{X^*}(\widetilde{F}'(x^k), F'(x^k)) - 4s^2 \le \varsigma_p^2 b_X^2(x, x^k) + e_{p,k}(x), \tag{2.9}$$

where, for $\psi_j := \alpha_u \kappa_u^{-j} \pi_u + \alpha_w[\iota_j \mu_u \pi_u + \kappa_w^{-j} \pi_w]$ and $\overline{\kappa} := \max\{\kappa_u, \kappa_w\}$, we set

$$\varsigma_p := \frac{\overline{\kappa}}{p} \sum_{j=0}^{\infty} p^j \psi_j \le \frac{(\alpha_u \pi_u + \alpha_w \pi_w)\kappa\overline{\kappa}}{p(\kappa - p)} + \frac{\alpha_w \mu_u \pi_u \overline{\kappa}}{p^2(\kappa - p)^2} \quad \text{and} \tag{2.10}$$

$$\begin{aligned} e_{p,k}(x) := {} & \frac{\varsigma_p(\alpha_u \kappa_u^{-k} + \alpha_w \iota_k \mu_u)}{\pi_u p^k} d_U^2(u^1, S_u(x^0)) + \frac{\varsigma_p \alpha_w \kappa_w^{-k}}{\pi_w p^k} d_W^2(w^1, S_w(x^0)) \\ & + \sum_{j=0}^{k-1} \frac{\varsigma_p \psi_{k-j}}{p^{k-j}} b_X^2(x^{j+1}, x^j) - \varsigma_p^2 b_X^2(x, x^k). \end{aligned} \tag{2.11}$$

Proof. By the differential transformation Assumption 2.2 (iii), we have

$$d_{X^*}(\widetilde{F}'(x^k), F'(x^k)) \le \alpha_u d_U(u^{k+1}, S_u(x^k)) + \alpha_w d_W(w^{k+1}, S_w(x^k)) =: R^{k+1}.$$

Since $\{x^n\}_{n=0}^k \subset \Omega$, the inner and adjoint tracking Assumption 2.2 (i) and (ii) give

$$\begin{aligned} \kappa_u d_U(u^{k+1}, S_u(x^k)) &\le d_U(u^k, S_u(x^{k-1})) + \pi_u b_X(x^k, x^{k-1}) \quad \text{and} \\ \kappa_w d_W(w^{k+1}, S_w(x^k)) &\le d_W(w^k, S_w(x^{k-1})) + \mu_u d_U(u^{k+1}, S_u(x^k)) + \pi_w b_X(x^k, x^{k-1}). \end{aligned}$$

Thus, invoking Lemma 2.10 with $b_{k+1} = d_U(u^{k+1}, S_u(x^k))$, $c_{k+1} = d_W(w^{k+1}, S_w(x^k))$, and $d_{k+1} = b_X(x^{k+1}, x^k)$, we obtain

$$R^{k+1} \le (\alpha_u \kappa_u^{-k} + \alpha_w \iota_k \mu_u) d_U(u^1, S_u(x^0)) + \alpha_w \kappa_w^{-k} d_W(w^1, S_w(x^0)) + \sum_{j=0}^{k-1} \psi_{k-j} b_X(x^{j+1}, x^j).$$

Using Young's inequality several times here, and adding the productive zero,

$$\begin{aligned} 4sR^{k+1} \le {} & \frac{(\alpha_u \kappa_u^{-k} + \alpha_w \iota_k \mu_u)^2}{\theta_k^u} d_U^2(u^1, S_u(x^0)) + \frac{(\alpha_w \kappa_w^{-k})^2}{\theta_k^w} d_W^2(w^1, S_w(x^0)) \\ & + \sum_{j=0}^{k-1} \frac{\psi_{k-j}^2}{\theta_{k,j}} b_X^2(x^{j+1}, x^j) + 4\left(\theta_k^u + \theta_k^w + \sum_{j=0}^{k-1} \theta_{k,j}\right) s^2 + \varsigma_p^2 b_X^2(x, x^k) - \varsigma_p^2 b_X^2(x, x^k) \end{aligned} \tag{2.12}$$

for any $\theta_k^u, \theta_k^w, \theta_{k,j} > 0$. Take $\theta_k^u = p^k \varsigma_p^{-1} \pi_u(\alpha_u \kappa_u^{-k} + \alpha_w \iota_k \mu_u)$, $\theta_k^w = p^k \varsigma_p^{-1} \pi_w \alpha_w \kappa_w^{-k}$, and $\theta_{k,j} = \varsigma_p^{-1} p^{k-j} \psi_{k-j}$. Observe from (2.6) that $\iota_k \le \kappa_w \iota_{k+1}$. Hence $p^k \iota_k \le (\kappa_w/p) p^{k+1} \iota_{k+1}$, and further, $p^k \psi_k \le (\overline{\kappa}/p) p^{k+1} \psi_{k+1}$, where $\overline{\kappa}/p > 1$. Now

$$\theta_k^u + \theta_k^w + \sum_{j=0}^{k-1} \theta_{k,j} = \frac{1}{\varsigma_p}\left(p^k \psi_k + \sum_{j=1}^{k} p^j \psi_j\right) \le \frac{\overline{\kappa}}{\varsigma_p p} \sum_{j=0}^{k+1} p^j \psi_j \le 1.$$

Thus rearranging (2.12) establishes (2.9). Finally, the bound in (2.10) on ς_p follows from (2.8) and $\sum_{j=0}^{\infty}(p/\kappa)^j = 1/(1-p/\kappa) = \kappa/(\kappa - p)$. □

Lemma 2.12. *Let $\{e_{p,k}(x^{k+1})\}_{k=0}^{N-1}$ be as in Lemma 2.11 with $p \geq 1$. Then*

$$\sum_{k=0}^{N-1} p^k e_{p,k}(x^{k+1}) \leq \frac{d_U^2(u^1, S_u(x^0))}{\pi_u}\left(\frac{\varsigma_p \alpha_u \kappa}{\kappa - 1} + \frac{\varsigma_p \alpha_w \mu_u}{(\kappa-1)^2}\right) + \frac{d_W^2(w^1, S_w(x^0))}{\pi_w}\left(\frac{\varsigma_p \alpha_w \kappa}{\kappa - 1}\right).$$

Proof. We split $p^k e_{p,k}(x^{k+1}) =: A_k + B_k + C_k - D_k$ observing (2.11). Now

$$\begin{aligned}\sum_{k=0}^{N-1} C_k = \sum_{k=0}^{N-1} p^k \sum_{j=0}^{k-1} \frac{\varsigma_p \psi_{k-j}}{p^{k-j}} b_X^2(x^{j+1}, x^j) &= \varsigma_p \sum_{j=0}^{N-2} p^j \sum_{k=j+1}^{N-1} \psi_{k-j} b_X^2(x^{j+1}, x^j) \\ &= \varsigma_p \sum_{j=0}^{N-2} p^j \sum_{\ell=0}^{N-2-j} \psi_{\ell+1} b_X^2(x^{j+1}, x^j) \leq \sum_{j=0}^{N-2} p^j \varsigma_p^2 b_X^2(x^{j+1}, x^j) \leq \sum_{k=0}^{N-1} D_k.\end{aligned}$$

Moreover, using (2.8) and the sum formula for geometric series, we estimate that $\sum_{k=0}^{N-1}(A_k + B_k)$ is less than the right-hand side of the claim. □

2.3 SMOOTHNESS OF DIFFERENTIAL ESTIMATES

We can now produce descent inequalities and Lipschitz estimates where $\widetilde{F'}(x^k)$ replaces $F'(x^k)$. Simpy taking $\bar{x} = x^k$ in the corollary to follow, and combining with the descent inequality

$$\langle F'(x^k) | x - x^k \rangle_{X^*,X} \geq F(x) - F(x^k) - \frac{L}{2} b_X^2(x, x^k),$$

produces a “descent inequality with error” for $\widetilde{F'}$. Likewise, we obtain a “three-point” descent inequality when we combine the result with (see [6, 27])

$$\langle F'(x^k) | x - \bar{x} \rangle_{X^*,X} \geq F(x) - F(\bar{x}) + \frac{\beta}{2} d_X^2(x, \bar{x}) - \frac{L}{2} b_X^2(x, x^k).$$

Assuming normed spaces, we say that d_{X^*} is *Young to* d_X, if both are one-homogeneous and

$$\langle \tilde{x}^* - x^* | \tilde{x} - x \rangle_{X^*,X} \leq \frac{1}{2} d_{X^*}^2(\tilde{x}^*, x^*) + \frac{1}{2} d_X^2(x^*, x) \quad \text{for all} \quad x, \tilde{x} \in X \text{ and } x^*, \tilde{x}^* \in X^*.$$

This obviously holds for d_X and d_{X^*} induced by the respective norms.

Corollary 2.13. *Suppose Assumption 2.2 holds, X is a normed space, X^* its dual with $d_{X^*}^2$ Young to d_X^2. Also let $\{x^n\}_{n=0}^k \subset \Omega$ for a $k \in \mathbb{N}$, and pick $p \in [1, \kappa)$. Then, for $e_{p,k}$ and ς_p defined in Lemma 2.11, we have $\sup_{N \in \mathbb{N}} \sum_{k=0}^{N-1} p^k e_{p,k}(x^{k+1}) < \infty$ and, for any $\tilde{\gamma} > 0$ and $x, \bar{x} \in X$,*

$$\langle \widetilde{F'}(x^k) - F'(x^k) | x - \bar{x} \rangle_{X^*,X} \geq -\frac{\tilde{\gamma}}{2} d_X^2(x, \bar{x}) - \frac{\varsigma_p^2}{2\tilde{\gamma}} b_X^2(x, x^k) - \frac{1}{2\tilde{\gamma}} e_{p,k}(x).$$

Proof. Take $s = d_{X^*}(\widetilde{F'}(x^k), F'(x^k))/2$ in Lemma 2.11. Then (2.9) reads

$$d_{X^*}^2(\widetilde{F'}(x^k), F'(x^k)) \leq \varsigma_p^2 b_X^2(x, x^k) + e_{p,k}(x^k).$$

By the Young relationship of d_X and d_X^*,

$$\langle \widetilde{F'}(x^k) - F'(x^k) | x - \bar{x} \rangle_{X^*,X} \geq -\frac{1}{2\tilde{\gamma}} d_{X^*}(\widetilde{F'}(x^k), F'(x^k)) - \frac{\tilde{\gamma}}{2} d_X^2(x, \bar{x}).$$

Together these two inequalities establish the claimed inequality. Lemma 2.12 shows the boundedness of $\sum_{k=0}^{N-1} p^k e_{p,k}(x^{k+1})$. □

Taking $x^* = F'(x)$ in the following theorem, and combining with F' being Lipschitz, we can get a Lipschitz-like property with error for $\widetilde{F'}$.

Theorem 2.14. *Suppose Assumption 2.2 holds, d_{X^*} satisfies the triangle inequality, and that $\{x^n\}_{n=0}^k \subset \Omega$ for a $k \in \mathbb{N}$. Then, for any $\vartheta > 0$,*

$$\frac{1}{2}d^2_{X^*}(\widetilde{F'}(x^k), x^*) \leq \frac{1+\vartheta}{2}d^2_{X^*}(F'(x^k), x^*) + \frac{1+\vartheta^{-1}}{2}e_{\mathrm{lip},k} \quad \text{for all} \quad x^* \in X^*, \tag{2.13}$$

where

$$\begin{aligned} e_{\mathrm{lip},k} &:= \frac{\varsigma_1}{\pi_u}(\alpha_u\kappa_u^{-k} + \alpha_w\iota_k\mu_u)d^2_U(u^1, S_u(x^0)) + \frac{\varsigma_1}{\pi_w}\alpha_w\kappa_w^{-k}d^2_W(w^1, S_w(x^0)) \\ &\quad + \sum_{j=0}^{k-1}\varsigma_1\psi_{k-j}d^2_X(x^{j+1}, x^j). \end{aligned} \tag{2.14}$$

Moreover, if $C > 0$ is a constant independent of N, then

$$\sup_{N\in\mathbb{N}}\sum_{k=0}^{N-1}e_{\mathrm{lip},k} < \infty \quad \textit{whenever} \quad \sum_{k=0}^{N-1}b^2_X(x^{j+1}, x^j) < C.$$

Proof. We apply Lemma 2.11 with $s = d_{X^*}(\widetilde{F'}(x^k), F'(x^k))/2$, $p = 1$, and $x = x^k$. With these choices $e_{1,k}(x^k) = e_{\mathrm{lip},k}$, and (2.9) reads

$$d^2_{X^*}(\widetilde{F'}(x^k), F'(x^k)) \leq e_{\mathrm{lip},k}.$$

With this, (2.13) follows after we use the triangle and Young's inequalities to derive

$$\frac{1}{2}d^2_{X^*}(\widetilde{F'}(x^k), x^*) \leq \frac{1+\vartheta}{2}d^2_{X^*}(F'(x^k), x^*) + \frac{1+\vartheta^{-1}}{2}d^2_{X^*}(\widetilde{F'}(x^k), F'(x^k)).$$

For the sum of $e_{\mathrm{lip},k}$, we start with the first two terms of (2.14). Using (2.8) and the sum formula for geometric series, we bound their sum over $k \in \{0, \ldots, N-1\}$ by

$$d^2_U(u^1, S_u(x^0))\frac{\varsigma_1}{\pi_u}\left(\frac{\alpha_u\kappa}{\kappa-1} + \frac{\alpha_w\mu_u}{(\kappa-1)^2}\right) + d^2_W(w^1, S_w(x^0))\frac{\varsigma_1}{\pi_w}\left(\frac{\alpha_w\kappa}{\kappa-1}\right).$$

We have $\varsigma_1 < \infty$ by Lemma 2.11. We sum the third term of (2.14) over $k \in \{0, \ldots, N-1\}$, and change the order of summation to obtain

$$\varsigma_1\sum_{k=0}^{N-1}\sum_{j=0}^{k-1}\psi_{k-j}b^2_X(x^{j+1}, x^j) = \varsigma_1\sum_{j=0}^{N-2}\left(\sum_{\ell=0}^{N-2-j}\psi_{\ell+1}\right)b^2_X(x^{j+1}, x^j) \leq \frac{\varsigma_1^2}{\kappa}\sum_{j=0}^{N-2}b^2_X(x^{j+1}, x^j).$$

Thus $\sup_{N\in\mathbb{N}}\sum_{k=0}^{N-1}e_{\mathrm{lip},k} < \infty$ whenever $\sup_{N\in\mathbb{N}}\sum_{j=0}^{N-1}b^2_X(x^{j+1}, x^j) < C$. ◻

3 OPERATOR-RELATIVE REGULARITY

To facilitate treating primal-dual methods as forward-backward methods with respect to suitable operators in the next section, we will now introduce operator-relative smoothness and monotonicity concepts.

3.1 DEFINITIONS

For a self-adjoint positive semi-definite $\Lambda \in \mathbb{L}(X;X^*)$ on a normed space X, we say that the Gâteaux derivative DF of $F : X \to \mathbb{R}$ is Λ-*firmly Lipschitz* if

$$\langle DF(z) - DF(x)|h\rangle_{X^*,X} \le \|x-z\|_\Lambda \|h\|_\Lambda \quad (x,z,h \in X).$$

This implies for any $h \in X$ with $\|h\|_X \le 1$ that

$$\|DF(z) - DF(x)\|^2_{X^*} \le \|z-x\|^2_\Lambda \|h\|^2_\Lambda \le \|\Lambda\|_{\mathbb{L}(X;X^*)}\|z-x\|^2_\Lambda \le \|\Lambda\|^2_{\mathbb{L}(X;X^*)}\|z-x\|^2_X. \tag{3.1}$$

In particular, DF is $\|\Lambda\|$-Lipschitz.

Likewise, we call DF locally Γ-monotone in $\Omega \ni \bar{x}$ for a self-adjoint $\Gamma \in \mathbb{L}(X;X^*)$ if

$$\langle DF(z) - DF(\bar{x})|z-\bar{x}\rangle \ge q_\Gamma(z-\bar{x}) \qquad \text{for} \quad q_\Gamma(x) := \langle \Gamma x|x\rangle_{X^*,X} \quad (z \in \Omega).$$

We do not at this stage assume Γ to be positive semi-definite. We, however, call Γ *Young* if there exists a self-adjoint positive semi-definite $|\Gamma| \in \mathbb{L}(X;X^*)$ such that

$$2\langle \Gamma x|z\rangle_{X^*,X} \le \|x\|^2_{|\Gamma|} + \|z\|^2_{|\Gamma|} \quad (x,z \in X).$$

Likewise, we call G Γ-subdifferentiable and ∂G Γ-monotone if, respectively,

$$G(\tilde{x}) - G(x) \ge \langle q|\tilde{x}-x\rangle + \frac{1}{2}q_\Gamma(\tilde{x}-x) \quad \text{or} \quad \langle \tilde{q}-q|\tilde{x}-x\rangle \ge q_\Gamma(\tilde{x}-x)$$

for all $q \in \partial G(x)$; $\tilde{q} \in \partial\tilde{G}(x)$, and $x, \tilde{x} \in X$.

Remark 3.1. Aside from $\|\cdot\|_M$, introduced in the next section, which presently needs to satisfy the Pythagoras' identity, our work does not strictly depend on $\|\cdot\|_\Lambda$, $\|\cdot\|_{|\Gamma|}$ or q_Γ being produced by operators. They could be arbitrary semi-norms and quadratics, if we introduced the formal calculus $q_{a\Lambda+b\Gamma} := a\|\cdot\|^2_\Lambda + bq_\Gamma$, etc., for $a, b \in \mathbb{R}$. For simplicity, we have chosen to work with operators.

3.2 ESTIMATES

We first prove a Λ-firmly Lipschitz descent lemma.

Lemma 3.2. *On a normed space X, suppose $F : X \to \mathbb{R}$ has a Λ-firmly Lipschitz Gâteaux derivative for a self-adjoint positive semi-definite $\Lambda \in \mathbb{L}(X;X^*)$. Then*

$$F(x) - F(z) - \langle DF(z)|x-z\rangle_{X^*,X} \le \frac{1}{2}\|z-x\|^2_\Lambda. \tag{3.2}$$

Proof. By the mean value theorem and the assumed firm Lipschitz property,

$$F(x) - F(z) - \langle DF(z)|x-z\rangle_{X^*,X} = \int_0^1 \langle DF(z+t(x-z)) - DF(z)|x-z\rangle \,dt \le \int_0^1 t\|x-z\|^2_\Lambda \,dt.$$

Integrating, the claim follows. □

We then prove a three-point smoothness lemma for nonconvex functions. Compared to [27, Appendix B], it is important that x ($= x^{k+1}$ in the application to forward steps at x^k) is not a priori restricted to the neighbourhood Ω of Γ-monotonicity at $\bar{x}$.

Lemma 3.3. *On a normed space X, let $F : X \to \mathbb{R}$ and suppose DF is Λ-Lipschitz for some $\Lambda \in \mathbb{L}(X;X^*)$, and Γ-monotone at $\bar{x} \in X$ in a neighbourhood $\Omega \ni \bar{x}$ and a Young $\Gamma \in \mathbb{L}(X;X^*)$. Then, for any $\beta > 0$, for all $z \in \Omega$ and $x \in X$,*

$$\langle DF(z)|x-\bar{x}\rangle \ge F(x) - F(\bar{x}) + \frac{1}{2}q_{\Gamma-\beta|\Gamma|}(x-\bar{x}) - \frac{1}{2}q_{\Lambda+\beta^{-1}|\Gamma|-\Gamma}(x-z).$$

Proof. Similarly to the proof of the descent inequality in Lemma 3.2, the mean value theorem applied to $\varphi(t) := F(\bar{x} + t(z - \bar{x}))$, followed by the assumed local Γ-monotonicity of DF, and the Young property of Γ, establishes

$$\begin{aligned} &F(\bar{x}) - F(z) - \langle DF(z)|\bar{x} - z\rangle_{X^*,X} \\ &= \int_0^1 \langle DF(z + t(\bar{x} - z)) - DF(z)|\bar{x} - z\rangle \,\mathrm{d}t \geq \int_0^1 t q_\Gamma(\bar{x} - z)\,\mathrm{d}t = \frac{1}{2} q_\Gamma(\bar{x} - z) \\ &= \frac{1}{2} q_\Gamma(x - \bar{x}) + \frac{1}{2} q_\Gamma(x - z) - \langle \Gamma(x - \bar{x})|x - z\rangle \geq \frac{1}{2} q_{\Gamma - \beta|\Gamma|}(x - \bar{x}) + \frac{1}{2} q_{\Gamma - \beta^{-1}|\Gamma|}(x - z). \end{aligned}$$

Applying Lemma 3.2 and summing this inequality with the descent inequality it provides, we obtain the claim. ◻

The next, monotonicity version of the previous lemma, slightly improves [6, Lemma 15.1] even in the Hilbert space scalar factor case.

Lemma 3.4. *On a normed space X, let $F : X \to \mathbb{R}$ and suppose DF is Λ-Lipschitz for some $\Lambda \in \mathbb{L}(X; X^*)$, and Γ-monotone for a self-adjoint $\Gamma \in \mathbb{L}(X; X^*)$ in a neighbourhood $\Omega \ni \bar{x}$ for some $\bar{x} \in X$. Then, for any $\beta, \zeta > 0$, for all $z \in \Omega$ and $x \in X$, with $\tilde{\Gamma} := \Gamma - (\zeta/2)\Lambda$, we have*

$$\langle DF(z) - DF(\bar{x})|x - \bar{x}\rangle_{X^*,X} \geq q_{\tilde{\Gamma} - \beta|\tilde{\Gamma}|} q(x - \bar{x}) - q_{\Lambda/(2\zeta) + \beta^{-1}|\tilde{\Gamma}| - \tilde{\Gamma}}(x - z).$$

Proof. Using both the Γ-monotonicity and the Λ-firmly Lipschitz property, and finishing with Young's inequality, we obtain

$$\begin{aligned} \langle DF(z) - DF(\bar{x})|x - \bar{x}\rangle_{X^*,X} &= \langle DF(z) - DF(\bar{x})|z - \bar{x}\rangle_{X^*,X} + \langle DF(z) - DF(\bar{x})|x - z\rangle_{X^*,X} \\ &\geq q_\Gamma(z - \bar{x}) - \|z - \bar{x}\|_\Lambda \|x - z\|_\Lambda \geq q_{\tilde{\Gamma}}(z - \bar{x}) - \frac{1}{2\zeta}\|x - z\|_\Lambda^2. \end{aligned}$$

Arguing for $q_{\tilde{\Gamma}}$ as in the proof of Lemma 3.3, we obtain the claim. ◻

4 NONCONVEX FORWARD-BACKWARD TYPE METHODS WITH INEXACT UPDATES

In this section, we work with abstract forward backward-type methods in a normed space X. The starting point is the problem

$$\min_{x \in X} F(x) + G(x),$$

where $F, G : X \to \overline{\mathbb{R}}$ are proper. In our general theory, we will directly make no further assumptions on the functions, although in this initial discussion and the examples of Section 4.2, G will be convex and lower semicontinuous, and F Fréchet differentiable.

For an initial x^0, if X is Hilbert, the iterates $\{x^k\}_{k=1}^\infty$ for the basic inexact forward-backward method are generated for some step length parameter $\tau > 0$ by

$$x^{k+1} := \operatorname{prox}_{\tau G}(x^k - \tau \widetilde{\nabla F}(x^k)), \tag{4.1}$$

where $\widetilde{\nabla F}(x^k)$ is an estimate of $\nabla F(x^k)$ (not necessarily one from Section 2). In implicit form the method reads

$$-\tau^{-1}(x^{k+1} - x^k) \in \widetilde{\nabla F}(x^k) + \partial G(x^{k+1}).$$

We generalise this method to saddle point problems by considering for a skew-adjoint $\Xi \in \mathbb{L}(X; X^*)$, i.e., $\Xi^*|X = -\Xi$, the problem of finding $x \in X$ satisfying

$$0 \in H(x) := F'(x) + \partial G(x) + \Xi x. \tag{4.2}$$

To allow, besides the inexact gradients of Section 2, inexact proximal maps [30], and mismatched adjoints [15], we consider the general inexact implicit algorithm

$$-M(x^{k+1}-x^k) =: \tilde{\partial}_{k+1} \mathrel{\tilde{\in}} F'(x^k) + \partial G(x^{k+1}) + \Xi x^{k+1}, \tag{4.3}$$

where "$\tilde{\in}$" stands for approximate inclusion (to be made more precise later), and the *preconditioning operator* $M \in \mathbb{L}(X; X^*)$ is self-adjoint and positive semi-definite. We could generalise M to a Bregman divergence, but choose simplicity of presentation; see, however, Remark 4.20.

Example 4.1 (Forward-backward splitting). For forward-backward splitting with inexact $\widetilde{F'}(x^k) \approx F'(x^k)$, we take $\tilde{\partial}_{k+1} \in \widetilde{F'}(x^k) + \partial G(x^{k+1})$ with $M = \tau^{-1}\operatorname{Id}$ and $\Xi = 0$.

Algorithms of the form (4.3) with exact inclusion for $\tilde{\partial}_{k+1}$ cover many common splitting algorithms, such as Douglas–Rachford splitting (DRS) and the primal-dual proximal splitting (PDPS) of [3]; see [6, 27]. With an inexact inclusion, they also cover the forward-backward method of [29] for point source localisation in measure spaces.

Example 4.2 (Primal-dual proximal splitting). On normed spaces Z and Y, let $g : Z \to \overline{\mathbb{R}}$ and $h : Y^* \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, $f : Z \to \mathbb{R}$ possibly non-convex but Fréchet differentiable, and $K \in \mathbb{L}(Z; Y^*)$. Suppose $h = (h^*)^*$ for some $h^* : Y \to \overline{\mathbb{R}}$, and consider the problem

$$\min_{z\in Z} f(z) + g(z) + h(Kz) = \min_{z\in Z}\max_{y\in Y} f(z) + g(z) + \langle y|Kz\rangle_{Y,Y^*} - h^*(y).$$

If f is convex, the Fenchel–Rockafellar theorem gives rise to the necessary and sufficient first-order primal-dual optimality conditions

$$0 \in H(z,y) = \begin{pmatrix} \partial g(z) + f'(z) + K^* y \\ \partial h^*(y) - Kz. \end{pmatrix} = F'(z,y) + \partial G(z,y) + \Xi(z,y),$$

where $F(z,y) = f(z)$, $G(z,y) = g(z) + h^*(y)$, and $\Xi = \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix}$. If f is nonconvex, the necessity can be shown through, e.g., Mordukhovich subdifferentials, and their compatibility with both convex subdifferentials and Fréchet derivatives; see, e.g., [6].

Pick step length parameters $\tau, \sigma > 0$. With inexact gradients for f, the PDPS in Hilbert spaces then reads

$$\begin{cases} z^{k+1} := \operatorname{prox}_{\tau g}(z^k - \tau\widetilde{\nabla f}(z^k) - \tau K^* y^k), \\ y^{k+1} := \operatorname{prox}_{\sigma h^*}(y^k - \sigma K(2z^{k+1} - z^k)). \end{cases} \tag{4.4}$$

When $f = j \circ S_u$ for S_u a PDE solution operator, and we compute $\widetilde{\nabla f}$ following Examples 2.5, 2.6 and 2.9, (4.4) becomes the algorithm presented in [14].

To extend (4.4) to general normed spaces, we write it in $X = Z \times Y$ in implicit form as (4.3) with $\tilde{\partial}_{k+1} \in \widetilde{F'}(x^k) + \partial G(x^{k+1}) + \Xi x^{k+1}$, where

$$\widetilde{F'}(z^k, y^k) := \begin{pmatrix} \widetilde{f'}(z^k) \\ 0 \end{pmatrix} \quad\text{and}\quad M := \begin{pmatrix} \tau^{-1}M_z, & K^* \\ K & \sigma^{-1}M_y \end{pmatrix}$$

for some self-adjoint positive semi-definite $M_z \in \mathbb{L}(Z; Z^*)$ and $M_y \in \mathbb{L}(Y; Y^*)$. For standard proximal maps in Hilbert spaces, $M_z = \operatorname{Id}$ and $M_y = \operatorname{Id}$. In that case, M is self-adjoint and positive semi-definite when $\tau\sigma\|K\|^2 \le 1$, while the treatment of exact forward steps with respect to f

requires[1] $\tau\lambda + \tau\sigma\|K\|^2 \le 1$ for λ the Lipschitz factor of f' [27, 6, 12]. In normed spaces, we extend this to the following, where in the standard Hilbert setting with $M_y = \mathrm{Id}$ and $M_z = \mathrm{Id}$, we can take $K_z = K$ and $K_y = \mathrm{Id}$.

Assumption 4.3 (PDPS step length condition). In the setting of Example 4.2, M_z is positive semi-definite, and $K = K_y K_z$ for some $K_z \in \mathbb{L}(Z;V)$, $K_y \in \mathbb{L}(V;Y^*)$, and a normed space V. Given $\lambda \ge 0$, the step length parameters $\tau, \sigma > 0$ satisfy

$$K_y K_y^* \le M_y \quad \text{and} \quad \tau\lambda M_z + \tau\sigma K_z^* K_z \le M_z.$$

Lemma 4.4 (PDPS preconditioning operator). *If Assumption 4.3 holds, then M is positive semi-definite and for any $\gamma_z, \gamma_y \ge 0$ and $\gamma := \min\{\gamma_z\tau, \gamma_y\sigma\}/2$, we have*

$$\lambda\,\mathrm{diag}(M_z,\ 0) \le M \quad \textit{and} \quad \gamma M \le \mathrm{diag}(\gamma_z M_z,\ \gamma_y M_y).$$

Proof. By a simple application of Young's inequality and Assumption 4.3, we have

$$\|(z,y)\|_M^2 = \tau^{-1}\|z\|_{M_z} + \sigma^{-1}\|y\|_{M_y} - 2\langle K_z z | K_y^* y\rangle_{Y^*,Y} \ge \|z\|^2_{\tau^{-1}M_z - \sigma K_z^* K_z} \ge \lambda\|z\|^2_{M_z}$$

for any $x = (z,y) \in Z \times Y$. This establishes the first claimed inequality. The second follows by using Young's inequality and Assumption 4.3 to establish

$$\gamma\|(z,y)\|_M^2 \le \gamma\|z\|^2_{\tau^{-1}M_z + \sigma K_z^* K_z} + \gamma\|y\|^2_{\sigma^{-1}M_y + \sigma^{-1}K_y^* K_y} \le \frac{2\gamma}{\tau}\|z\|^2_{M_z} + \frac{2\gamma}{\sigma}\|y\|^2_{M_y}.$$

□

Remark 4.5 (Testing operators and accelerated methods). In [27, 6], "testing operators" $Z_k \in \mathbb{L}(X^*;X^*)$ are used to encode convergence rates, and to prove "accelerated" $O(1/N^2)$ rates for the PDPS under mere primal strong convexity. They could be incorporated into our treatise, however, for simplicity, we have chosen not to do this.

4.1 INEXACT GROWTH INEQUALITIES

We now make precise the approximate inclusion in (4.3). We define the *Lagrangian gap functional*

$$\mathcal{G}(x;\bar{x}) := [F+G](x) - [F+G](\bar{x}) - \langle \Xi x | \bar{x}\rangle_{X^*,X}.$$

Example 4.6. For forward-backward splitting, $\mathcal{G}(x;\bar{x}) = [F+G](x) - [F+G](\bar{x})$ is simply a function value difference.

Example 4.7. For the PDPS of Example 4.2, with $x = (y,z)$, we expand

$$\mathcal{G}(x;\bar{x}) = \mathcal{L}(z,\bar{y}) - \mathcal{L}(\bar{z},y) \quad \text{for} \quad \mathcal{L}(z,y) := [f+g](z) + \langle Kz|y\rangle - h^*(y).$$

This is different from the true duality gap that arises from the Fenchel–Rockafellar theorem. For the latter no convergence results exist to our knowledge. In the convex case, if $0 \in H(\bar{x})$, the Lagrangian gap is non-negative, however, it may be zero even if $0 \notin H(\bar{x})$, unlike for the true duality gap.

[1]This is the requirement for gap estimates; for iterate estimates $\lambda/2$ in place of λ is sufficient. In [32] an overall factor 4/3 improvement is shown through an analysis that involves historical iterates.

For subdifferential convergence, we will need an inexact descent inequality:

Assumption 4.8. $M \in \mathbb{L}(X; X^*)$ is self-adjoint and positive semi-definite. Also,

(i) For a set $\Omega \subset X$, $\eta > 0$, and $\mathbb{L}(X; X^*) \ni \breve{\Lambda} \le 2(1-\eta)M$, whenever $\{x^n\}_{n=0}^k \subset \Omega$, for some errors $\varepsilon_{\text{desc},k} \in \mathbb{R}$, for any $k \in \mathbb{N}$, we have

$$\langle \tilde{\partial}_{k+1} | x^{k+1} - x^k \rangle_{X^*,X} \ge \mathcal{G}(x^{k+1}; x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \varepsilon_{\text{desc},k}. \tag{4.5}$$

(ii) The errors satisfy $r_{\text{desc}} := \sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} \varepsilon_{\text{desc},k} < \infty$.

(iii) We have $x^0 \in \Omega$, and for any $N \ge 1$, $\sum_{k=0}^{N-1} \mathcal{G}(x^{k+1}; x^k) \le r_{\text{desc}}$ implies $x^N \in \Omega$.

Remark 4.9. If $\Omega = X$, convergence will be global. In the examples of Section 2.1, $\Omega \ne X$ may arise from S_u, G, or J being only locally Lipschitz continuously differentiable.

We will also need the approximations $\tilde{\partial}_{k+1}$ to become better as the distance between the iterates shrinks, in the sense of

Assumption 4.10. For H defined in (4.2), we have

$$\sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} \|x^{k+1} - x^k\|^2_M < \infty \implies \lim_{k\to\infty} \inf_{x^*_{k+1}\in H(x^{k+1})} \|x^*_{k+1} - \tilde{\partial}_{k+1}\|^2_{X^*} = 0.$$

This can be proved through Lipschitz differential estimates, as we discuss below.

For function value and iterate convergence, we cannot work with just the iterates: we need to assume properties with respect to a base point $\bar{x} \in X$, usually a solution. For iterate convergence, we assume the three-point monotonicity type estimate

$$\langle \tilde{\partial}_{k+1} - H(\bar{x}) | x^{k+1} - \bar{x} \rangle_{X^*,X} \ge \gamma\|x^{k+1} - \bar{x}\|^2_M - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \varepsilon_k(\bar{x}), \tag{4.6}$$

for all $k \in \mathbb{N}$, *whenever* $\{x^n\}_{n=0}^k \subset \Omega_{\bar{x}}$ for an open neighbourhood $\Omega_{\bar{x}}$ of $\bar{x}$, a positive semi-definite self-adjoint $\breve{\Lambda} \in \mathbb{L}(X; X^*)$, errors $\varepsilon_k(\bar{x}) \in \mathbb{R}$, and a $\gamma \ge 0$.

For function value convergence, we need again a descent inequality similar to (4.5), now instantiated at the base point $\bar{x}$ instead of x^k. That is, for all $k \in \mathbb{N}$, we assume for some errors $\varepsilon_k(\bar{x}) \in \mathbb{R}$ *whenever* $\{x^n\}_{n=0}^k \subset \Omega_{\bar{x}}$ that

$$\langle \tilde{\partial}_{k+1} | x^{k+1} - \bar{x} \rangle_{X^*,X} \ge \mathcal{G}(x^{k+1}; \bar{x}) + \frac{\gamma}{2}\|x^{k+1} - \bar{x}\|^2_M - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \varepsilon_k(\bar{x}). \tag{4.7}$$

We write $\varepsilon_{\text{desc},k}(\bar{x}) := \varepsilon_k(\bar{x})$ when we need draw a distinction to (4.6).

The errors will also need to have a finite sum:

Assumption 4.11. Given $\bar{x} \in X$, for some $\eta, \gamma \ge 0$ and $0 \le \breve{\Lambda} \le (1-\eta)M$, either

(a) (4.6) holds, and $\bar{x} \in H^{-1}(0)$; or

(b) (4.7) holds, and $\inf_{x\in\Omega_{\bar{x}}} \mathcal{G}(x; \bar{x}) \ge 0$.

Moreover, $x^0 \in \mathbb{O}_M(\bar{x}, \sqrt{\delta^2 - 2r_p})$ and $\mathbb{O}_M(\bar{x}, \delta) \subset \Omega_{\bar{x}}$ for some $\delta > 0$ and $p \ge 1$ with

$$\frac{1}{2}\delta^2 > r_p := \sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} p^{k-N}\varepsilon_k(\bar{x}) < \infty \quad \text{and} \quad p \le \overline{p}_\gamma := \begin{cases} 1 + 2\gamma & \text{in option (a),} \\ 1 + \gamma & \text{in option (b).} \end{cases} \tag{4.8}$$

4.2 EXAMPLES

We first take in Example 4.1 exact F' as well as $\widetilde{F'}$ based on Section 2. Then we consider variants of the PDPS of Example 4.2. In the proofs of Assumption 4.8 below, only its part (iii) requires $\Xi = 0$.

Theorem 4.12 (Operator-relative exact forward-backward splitting). *On a normed space X, let $F : X \to \mathbb{R}$ have a Λ-firmly Lipschitz Fréchet derivative for a self-adjoint positive semi-definite $\Lambda \in \mathbb{L}(X; X^*)$, and $G : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous. In (4.3), ensure $\tilde{\partial}_{k+1} \in F'(x^k) + \partial G(x^{k+1}) + \Xi x^{k+1}$. Then,*

(i) Assumption 4.8 holds with $\Omega = X$, $\varepsilon_{\mathrm{desc},k} = 0$ and $\check{\Lambda} = \Lambda$, provided $\Xi = 0$ and $\Lambda \le 2(1-\eta)M$ for an $\eta > 0$.

(ii) Assumption 4.10 holds if $\Lambda \le cM$ for a $c > 0$.

Suppose further that G is Γ_G-subdifferentiable, and F' is Γ_F-monotone in $\Omega_{\bar{x}} \supset \mathbb{O}_M(\bar{x}, \delta) \ni x_0$ for some $\Gamma_F, \Gamma_G \in \mathbb{L}(X; X^*)$, $\bar{x} \in X$ and $\delta > 0$. Let $\varepsilon_k(\bar{x}) := 0$. Then, for any $\gamma, \eta \ge 0$; as well as $p \in [1, \overline{p}_\gamma]$ ($p \in (0, \overline{p}_\gamma]$ if $\delta = \infty$):

(iii) Assumption 4.11 option (a) holds if, for $\tilde{\Gamma}_F := \Gamma_F - (\zeta/2)\Lambda$ and $\zeta, \beta > 0$,

$$\gamma M \le \Gamma_G + \tilde{\Gamma}_F - \beta|\tilde{\Gamma}_F| \quad \text{and} \quad \check{\Lambda} := \zeta^{-1}\Lambda + 2(\beta^{-1}|\tilde{\Gamma}_F| - \tilde{\Gamma}_F) \le (1-\eta)M.$$

(iv) Assumption 4.11 option (b) holds if, for $\beta > 0$,

$$\gamma M \le \Gamma_G + \Gamma_F - \beta|\Gamma_F| \quad \text{and} \quad \check{\Lambda} := \Lambda + \beta^{-1}|\Gamma_F| - \Gamma_F \le (1-\eta)M.$$

Proof. (i): By Lemma 3.2, we have

$$\langle F'(x^k) | x^{k+1} - x^k \rangle_{X^*,X} \ge F(x^{k+1}) - F(x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_\Lambda. \tag{4.9}$$

Since $\mathcal{G}(x^{k+1}; x^k) = [F+G](x^{k+1}) - [F+G](x^k)$ when $\Xi = 0$, combining (4.9) with the subdifferentiability of G verifies (4.5) with $\Omega = X$. Assumption 4.8 (i) ensues as we have assumed $\check{\Lambda} \le 2(1-\eta)M$. Our choices of Ω , $\varepsilon_{\mathrm{desc},k}$, and Ξ guarantee (ii) and (iii).

(ii): Since F is Λ-firmly Lipschitz, taking

$$x^*_{k+1} = F'(x^{k+1}) - F'(x^k) + \tilde{\partial}_{k+1} \in F'(x^{k+1}) + \partial G(x^{k+1}) + \Xi x^{k+1} = H(x^{k+1}) \tag{4.10}$$

and then using (3.1) and the assumption $\Lambda \le cM$, we estimate

$$\inf_{x^*_{k+1} \in H(x^{k+1})} \|x^*_{k+1} - \tilde{\partial}_{k+1}\|^2_{X^*} \le \|F'(x^{k+1}) - F'(x^k)\|^2_{X^*} \le c\|\Lambda\| \|x^{k+1} - x^k\|^2_M.$$

Thus the antecedent of Assumption 4.10 implies its consequent.

For the verification of both (iii) and (iv), we observe that (4.8) holds because $r_p = 0$ due to $\varepsilon_k(\bar{x}) = 0$. We have explicitly assumed the remaining conditions of Assumption 4.11, so we only need to verify the respective (4.6) or (4.7).

(iii): By Lemma 3.4 and our assumption $\check{\Lambda} \ge 0$, whenever $x^k \subset \Omega_{\bar{x}}$, we have

$$\langle F'(x^k) - F'(\bar{x}) | x^{k+1} - \bar{x} \rangle_{X^*,X} \ge q_{\tilde{\Gamma}_F - \beta\tilde{\Gamma}_F}(x^{k+1} - \bar{x}) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\zeta^{-1}\Lambda + 2(\beta^{-1}|\tilde{\Gamma}_F| - \tilde{\Gamma}_F)}. \tag{4.11}$$

Combining with the Γ_G-monotonicity of ∂G, the skew-symmetricity of Ξ, and the bound $\gamma M \le \Gamma_G + \tilde{\Gamma}_F - \beta|\tilde{\Gamma}_F|$, we verify (4.6).

(iv): By Lemma 3.3, whenever $x^k \subset \Omega_{\bar{x}}$, we have

$$\langle F'(x^k)|x^{k+1} - \bar{x}\rangle_{X^*,X} \geq F(x^{k+1}) - F(\bar{x}) + \frac{1}{2}q_{\Gamma_F - \beta|\Gamma_F|}(x^{k+1} - \bar{x}) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\Lambda+\beta^{-1}|\Gamma_F|-\Gamma_F}. \tag{4.12}$$

Combining with the Γ_G-subdifferentiability of G, the skew-symmetricity of Ξ, and the bound $\gamma M \leq \Gamma_G + \Gamma_F - \beta|\Gamma_F|$, we verify (4.7). □

Example 4.13 (Standard forward-backward on a Hilbert space). Let X be a Hilbert space, and suppose F' is L-Lipschitz for some $L \geq 0$, Take $M = \tau^{-1}\,\mathrm{Id}$ and $\Lambda = L\,\mathrm{Id}$ for a step length parameter $\tau > 0$. Then the condition $\Lambda \leq 2(1-\eta)M$ with $\eta > 0$ in Theorem 4.12 (i) reduces to the standard step length condition $\tau L < 2$.

With G γ_G-strongly subdifferentiable for some $\gamma_G \geq 0$, and F' γ_F-monotone for some $\gamma_F \in \mathbb{R}$, taking $\Gamma_G = \gamma_G\,\mathrm{Id}$, $\Gamma_F = \gamma_F\,\mathrm{Id}$, and $\check{\Lambda} = \check{\lambda}\,\mathrm{Id}$ for some $\check{\lambda} \geq 0$, the conditions in (iv) reduce to finding $\beta > 0$ such that

$$0 \leq \tau[L + \beta^{-1}|\gamma_F| - \gamma_F] \leq 1 \quad \text{and} \quad 0 \leq \gamma := \tau[\gamma_G + \gamma_F - \beta|\gamma_F|],$$

the first strictly for $\eta > 0$. The conditions of (iii) are analogous.

For the following, we recall that ς_p is defined in (2.10) and κ in (2.8), while r_{desc} and r_p are defined in Assumptions 4.8 and 4.11. We will take

$$d_X(x,\tilde{x}) = \|x - \tilde{x}\|_M, \quad d_{X^*}(x^*,\tilde{x}^*) = [2(\tfrac{1}{2}\|\,\cdot\,\|^2_M)^*(x^* - \tilde{x}^*)]^{1/2}, \quad b_X(x,\tilde{x}) = \|x - \tilde{x}\|_\Lambda. \tag{4.13}$$

Then the Fenchel–Young inequality and homogeneity ensure that d^*_X is Young to d_X, as defined in Section 2.3. If M is invertible, this gives $d_{X^*}(x^*,\tilde{x}^*) = \|x - \tilde{x}\|_{M^{-1}}$.

Theorem 4.14 (Forward-backward with inexact differentials). *On a normed space X, for a self-adjoint and positive semi-definite $\Lambda \in \mathbb{L}(X;X^*)$, suppose $F : X \to \mathbb{R}$ has a Λ-firmly Lipschitz Fréchet derivative, and $G : X \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous. For all $k \in \mathbb{N}$, construct $\widetilde{F'}(x^k)$ obeying Assumption 2.2 for the distances (4.13) and an $\Omega \subset X$. In (4.3), ensure $\tilde{\partial}_{k+1} \in \widetilde{F'}(x^k) + \partial G(x^{k+1}) + \Xi x^{k+1}$. Then:*

(i) Assumption 4.8 holds for any $\eta, \tilde{\gamma} > 0$, $p \in [1,\kappa)$, and $\varepsilon_{\mathrm{desc},k} = e_{p,k}(x^{k+1})/(2\tilde{\gamma})$ provided $\Xi = 0$, $\Omega \supset \mathrm{sub}_{r_{\mathrm{desc}}+[F+G](x^0)}(F+G)$, and

$$0 \leq \check{\Lambda} := (1 + \tilde{\gamma}^{-1}\varsigma_p)\Lambda + \tilde{\gamma}M \leq 2(1-\eta)M.$$

(ii) Assumption 4.10 holds if $\Lambda \leq cM$ for a $c > 0$.

Suppose further that G is Γ_G-strongly subdifferentiable, and F' is Γ_F-monotone in $\Omega_{\bar{x}} = \Omega \supset \mathbb{O}_M(\bar{x},\delta)$ for an $\bar{x} \in X$, $\delta > 0$, and $\Gamma_F, \Gamma_G \in \mathbb{L}(X;X^*)$. Pick $\tilde{\gamma} > 0$ and $p \in [1,\kappa)$. Set $\varepsilon_k(\bar{x}) = e_{p,k}(x^{k+1})/(2\tilde{\gamma})$ and define r_p by (4.8). If $x^0 \in \mathbb{O}_M(\bar{x},\sqrt{\delta^2 - 2r_p})$ with $r_p < \delta^2/2$, then, for any $\eta \geq 0$:

(iii) Assumption 4.11 option (a) holds if, for $\tilde{\Gamma}_F := \Gamma_F - (\zeta/2)\Lambda$ and $\zeta, \beta > 0$,

$$(\gamma + \tilde{\gamma})M \leq \Gamma_G + \tilde{\Gamma}_F - \beta|\tilde{\Gamma}_F| \quad \text{for a} \quad \gamma \geq (p-1)/2 \quad \text{and}$$
$$0 \leq \check{\Lambda} := (\zeta^{-1} + \varsigma_p\tilde{\gamma}^{-1})\Lambda + 2(\beta^{-1}|\tilde{\Gamma}_F| - \tilde{\Gamma}_F) \leq (1-\eta)M.$$

(iv) *Assumption 4.11 option (b) holds if, for $\beta > 0$,*

$$(\gamma + \tilde{\gamma})M \leq \Gamma_G + \Gamma_F - \beta|\Gamma_F| \quad \textit{for a} \quad \gamma \geq p - 1 \quad \textit{and}$$
$$0 \leq \breve{\Lambda} := (1 + \varsigma_p \tilde{\gamma}^{-1})\Lambda + \beta^{-1}|\Gamma_F| - \Gamma_F \leq (1 - \eta)M.$$

Proof. We first observe that, as in the proof of Theorem 4.12, F satisfies (4.9), (4.11) and (4.12) by Lemmas 3.2 to 3.4.

(i): Combining (4.9) with Corollary 2.13 for $\bar{x} = x^k$ and $x = x^{k+1}$ establishes

$$\langle \widetilde{F'}(x^k) | x^{k+1} - x^k \rangle_{X^*,X} \geq F(x^{k+1}) - F(x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \frac{1}{2\tilde{\gamma}} e_{p,k}(x^{k+1})$$

with $\sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} p^k e_{p,k}(x^{k+1}) < \infty$ whenever $\{x^n\}_{n=0}^k \subset \Omega$. Further combining with the subdifferentiability of G, we verify (4.5). Since we assume $\breve{\Lambda} \leq 2(1-\eta)M$ and take $\varepsilon_{\text{desc},k} \propto e_{p,k}(x^{k+1})$, this verifies Assumption 4.8 (i) and (ii). Because $\Xi = 0$, (iii) requires $[F + G](x^N) \leq r_{\text{desc}} + [F + G](x^0)$ to imply $x^N \in \Omega$. This holds whenever $\Omega \supset \operatorname{sub}_{r_{\text{desc}}+[F+G](x^0)}(F + G)$, as we have assumed.

(ii): Theorem 2.14 with $x^* = F'(x^{k+1})$ and $\vartheta = 1$ establishes

$$\sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} \|x^{k+1} - x^k\|^2_M \leq C \implies \sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} e_{\text{lip},k} < \infty \tag{4.14}$$

and, since F is Λ-firmly Lipschitz, together with (3.1) and $\Lambda \leq cM$,

$$\frac{1}{2}\|\widetilde{F'}(x^k) - F'(x^{k+1})\|^2_{X^*} \leq c\|\Lambda\|\|x^{k+1} - x^k\|^2_\Lambda + e_{\text{lip},k}. \tag{4.15}$$

The antecedent of (4.14) implies $\|\widetilde{F'}(x^k) - F'(x^{k+1})\|_{X^*} \to 0$ via (4.15). By choosing x^*_{k+1} of (4.10) in Assumption 4.10, this readily verifies the assumption.

For the verification of both (iii) and (iv), we observe that (4.8) holds because the lower bound on γ guarantees $p \in [1, \overline{p}_\gamma)$, and we have explicitly assumed $\delta^2 > 2r_p$, where $r_p = \sup_{N\in\mathbb{N}} p^{-N} \sum_{k=0}^{N-1} p^k e_{p,k}(x^{k+1})/(2\tilde{\gamma}) < \infty$ by Corollary 2.13. We have also explicitly assumed the remaining conditions of Assumption 4.11, so, only need to verify the respective (4.6) or (4.7).

(iii): Whenever $\{x^n\}_{n=0}^k \subset \Omega_{\bar{x}}$, combining (4.11) with Corollary 2.13 gives

$$\langle \widetilde{F'}(x^k) - F'(\bar{x}) | x^{k+1} - \bar{x} \rangle_{X^*,X} \geq q_{\tilde{\Gamma}_F - \beta\tilde{\Gamma}_F - \tilde{\gamma}M}(x^{k+1} - \bar{x}) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \frac{1}{2\tilde{\gamma}} e_{p,k}(x^{k+1}).$$

We verify (4.6) by combining this with the Γ_G-monotonicity of G and the skew-symmetricity of Ξ.

(iv): Whenever $\{x^n\}_{n=0}^k \subset \Omega_{\bar{x}}$, combining (4.12) with Corollary 2.13 gives

$$\begin{aligned}\langle \widetilde{F'}(x^k) | x^{k+1} - \bar{x} \rangle_{X^*,X} \geq{}& F(x^{k+1}) - F(\bar{x}) \\ &+ \frac{1}{2} q_{\Gamma_F - \beta|\Gamma_F| - \tilde{\gamma}M}(x^{k+1} - \bar{x}) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\breve{\Lambda}} - \frac{1}{2\tilde{\gamma}} e_{p,k}(x^{k+1}).\end{aligned}$$

We verify (4.7) by combining this with the (Γ_G-strong) subdifferentiability of G and the skew-symmetricity of Ξ. ◻

Example 4.15. Continuing from Example 4.13, in the standard scalar Hilbert space setting, the condition in (i) reduces to $0 \leq \tilde{\gamma} + \tau(1 + \varsigma_p \tilde{\gamma}^{-1})L < 2$. The conditions in (iv) reduce the step length $\tau > 0$ satisfying for some $\beta, \tilde{\gamma} > 0$ the bounds

$$0 \leq \tau[(1 + \varsigma_p \tilde{\gamma}^{-1})L + 2(\beta^{-1}|\gamma_F| - \gamma_F)] \leq 1 \quad \text{and} \quad 0 \leq \gamma := \tau[\gamma_G + \gamma_F - \beta|\gamma_F|] - \tilde{\gamma},$$

the first strictly for $\eta > 0$. Similarly we can translate (iii). Thus, we can expect the subdifferential convergence Assumption 4.8 to hold for small enough $\tau > 0$, but stronger forms of convergence via Assumption 4.11 will require the differential approximation to be good enough that $\varsigma_p L$ is small with respect to the available strong monotonicity. Studying (2.10), the main constants that control this quantity are π_u and π_w, which can always be made small if, instead a single iteration, we take sufficiently many iterations of the inner and adjoint solvers that satisfy Assumption 2.2 (i) and (ii).

Theorem 4.16 (PDPS with inexact $\widetilde{f'}$; everything else exact). *Assume the setup of Example 4.2 with g and h^* convex, and f with an LM_z-firmly Lipschitz Fréchet derivative for a $L \geq 0$. Suppose that Assumption 2.2 holds for f in $\Omega \subset Z$ with*

$$d_Z(z,\tilde z) = \|z-\tilde z\|_{M_z},\ d_{Z^*}(z^*,\tilde z^*) = [2(\tfrac{1}{2}\|\cdot\|^2_{M_z})^*(z^*-\tilde z^*)]^{1/2},\ \text{and}\ b_Z(z,\tilde z) = \|x-\tilde x\|_{LM_z}.$$

Ensure that the step length Assumption 4.3 holds for some $\tau,\sigma,\lambda > 0$. Then

(i) Assumption 4.10 holds.

Suppose further that g and h^* are, respectively, $\gamma_g M_z$ and $\gamma_{h^*} M_y$(-strongly) subdifferentiable for some $\gamma_g, \gamma_{h^*} \geq 0$, and that f' is γ_f-monotone in $\Omega_{\bar z} := \Omega \ni \bar z$. Let $\bar x \in \{\bar z\} \times \operatorname{dom} h^*$ and $\Omega_{\bar x} := \Omega_{\bar z} \times \operatorname{dom} h^*$. Suppose $\mathbb{O}_{M_z}(\bar z, \delta_z) \subset \Omega_{\bar z}$ for some $\delta_z > 0$. Pick $\tilde\gamma > 0$ and $p \in [1,\kappa)$. Set $\varepsilon_k(\bar x) = e_{p,k}(z^{k+1})/(2\lambda\tilde\gamma)$, and define r_p by (4.8). If

$$x^0 = (z^0, y^0) \in \mathbb{O}_M(\bar x, \sqrt{\lambda^2\delta_z^2 - 2r_p}) \quad \text{with} \quad \lambda^2\delta_z^2 > 2r_p, \tag{4.16}$$

then, for all $\eta \geq 0$:

(ii) Assumption 4.11 option (a) holds if, for $\tilde\gamma_f := \gamma_f - (\zeta/2)L$ and $\beta,\zeta > 0$,

$$\begin{aligned}(p-1)/2 \leq \gamma := \min\{(\gamma_g + \tilde\gamma_f - \beta|\tilde\gamma_f|)\tau, \gamma_{h^*}\sigma\}/2 - \tilde\gamma \quad \text{and}\\ 0 \leq \breve\lambda := \zeta^{-1}L + 2(\beta^{-1}|\tilde\gamma_f| - \tilde\gamma_f) + 2\varsigma_p\tilde\gamma^{-1}L \leq (1-\eta)\lambda.\end{aligned}$$

(iii) Assumption 4.11 option (b) holds if, for $\beta > 0$,

$$p - 1 \leq \gamma := \min\{(\gamma_g + \gamma_f - \beta|\gamma_f|)\tau, \gamma_{h^*}\sigma\}/2 - \tilde\gamma \quad \text{and} \tag{4.17}$$

$$0 \leq \breve\lambda := L + \beta^{-1}|\gamma_f| - \gamma_f + \varsigma_p\tilde\gamma^{-1}L \leq (1-\eta)\lambda. \tag{4.18}$$

Proof. F' is Λ-firmly Lipschitz and Γ_F-monotone, and G is Γ_G-strongly convex for

$$\Lambda := \operatorname{diag}(LM_z,\ 0), \quad \Gamma_F := \operatorname{diag}(\gamma_F M_z,\ 0), \quad \text{and} \quad \Gamma_G := \operatorname{diag}(\gamma_G M_z,\ \gamma_{h^*} M_y).$$

Adopting the distances (4.13) for X and X^*, Lemma 4.4 shows that

$$(\tfrac{1}{2}\|\cdot\|^2_M)^*((z^*,0)) = \sup_{(z,y)} \langle z^*|z\rangle - \frac{1}{2}\|(z,y)\|^2_M \leq \sup_z \langle z^*|z\rangle - \frac{\lambda}{2}\|z\|^2_{M_z} = \frac{1}{\lambda}(\tfrac{1}{2}\|\cdot\|^2_{M_z})^*(z^*).$$

Combining this with Assumption 2.2 for f and $\widetilde{f'}$ in $\Omega \subset Z$, we see that Assumption 2.2 holds in $\Omega \times \operatorname{dom} h^* \subset X$ for F and $\widetilde{F'}$ defined in Example 4.2 with α_u and α_w divided by λ compared to the case of f and $\widetilde{f'}$. This has the effect of dividing $e_{p,k}$ by λ. Our claims thus follow if we prove the remaining assumptions of Theorem 4.14.

(i): Lemma 4.4 proves $\Lambda \leq (L/\lambda)M$. Clearly $\Lambda \geq 0$. Now we use Theorem 4.14 (ii).

(iii): Observe that $|\Gamma_F| = \operatorname{diag}(|\gamma_f|M_z,\ 0)$. Taking $\gamma_z := \gamma_g + \gamma_f - \beta|\gamma_f|$ and $\gamma_y := \gamma_{h^*}$ in Lemma 4.4, and using (4.17) and (4.18), we obtain the required bounds

$$
\begin{aligned}
(1-\eta)M &\geq (1-\eta)\lambda \operatorname{diag}(M_z,\ 0) \geq \operatorname{diag}(\check\lambda M_z, 0) \\
&= \check\Lambda := \Lambda + \beta^{-1}|\Gamma_F| - \Gamma_F + \varsigma_p\tilde\gamma^{-1}\Lambda \quad \text{and} \\
\Gamma_G + \Gamma_F - \beta|\Gamma_F| &= \operatorname{diag}(\gamma_z M_z,\ \gamma_y M_y) \geq (1/2)\min\{\gamma_z\tau, \gamma_y\sigma\}M \geq (\gamma+\tilde\gamma)M.
\end{aligned}
\tag{4.19}
$$

Taking $\delta := \lambda\delta_z$, (4.16) implies, as required, $x^0 \in \mathbb{O}_M(\bar x, \sqrt{\delta^2 - 2r_p})$ and $2r_p < \delta^2$. By (4.19), we have $\mathbb{O}_M(\bar x, \delta) \subset \mathbb{O}_{M_z}(\bar z, \delta_z) \times \operatorname{dom} h^* \subset \Omega_{\bar z} \times \operatorname{dom} h^* = \Omega_{\bar x}$. By construction and assumption, we have $\check\Lambda \geq 0$. The claim now follows from Theorem 4.14 (iv).

(ii): completely analogous to (iii), observing that $\tilde\Gamma_F = \operatorname{diag}(\tilde\gamma_f M_z, 0)$. □

We finally consider adjoint mismatch as in [15], keeping everything else exact.

Theorem 4.17 (PDPS with adjoint mismatch). *Assume the setup of Example 4.2 with $\tau\sigma\|K\|^2 \leq 1$ and, for simplicity, $f = 0$ and Hilbert Z and Y. Suppose $\operatorname{dom} h^*$ is bounded, and that g and h^* are, respectively, γ_g- and γ_{h^*}-strongly convex for some $\gamma_g > 0$ and $\gamma_{h^*} \geq 0$. Let $\gamma := \min\{\gamma_g\tau/4, \gamma_{h^*}\sigma/2\}$. In the PDPS (4.4), not able to compute K^*, replace it with a "mismatched" adjoint $K^{*\approx}$. Then, for any $\bar x \in Z \times Y$ and $p \in (1, 1+2\gamma]$, Assumption 4.11 (a) holds with $\check\Lambda = 0$, $\Omega_{\bar x} = Z \times Y$, $\delta = \infty$, $r_p \leq \varepsilon/(1-p)$, and*

$$
\varepsilon_k(\bar x) = \frac{1}{2\gamma_g}\|(K^{*\approx} - K^*)y^k\|_Z^2 \leq \varepsilon := \frac{1}{2\gamma_g}(\|K^{*\approx} - K^*\| \operatorname{diam} \operatorname{dom} h^*)^2.
$$

Proof. With M, G, and F given by Example 4.2, the abstract algorithm (4.3) reads

$$
-M(x^{k+1} - x^k) =: \tilde\partial_{k+1} = x^*_{k+1} + ((K^{*\approx} - K^*)y^k, 0) \quad \text{for a} \quad x^*_{k+1} \in H(x^{k+1}).
$$

Here H is defined in (4.2). Using Lemma 4.4 in the final step, we estimate

$$
\begin{aligned}
\langle \tilde\partial_{k+1} - H(\hat x) | x^{k+1} - \bar x\rangle_{X^*,X} &= \langle \tilde\partial_{k+1} - x^*_{k+1} | x^{k+1} - \bar x\rangle_{X^*,X} + \langle x^*_{k+1} - H(\hat x) | x^{k+1} - \bar x\rangle_{X^*,X} \\
&\geq \langle (K^{*\approx} - K^*)y^k, z^{k+1} - \bar z\rangle + \gamma_g\|z^{k+1} - \bar z\|_Z^2 + \gamma_{h^*}\|y^{k+1} - \bar y\|_Y^2 \\
&\geq \frac{\gamma_g}{2}\|z^{k+1} - \bar z\|_Z^2 + \gamma_{h^*}\|y^{k+1} - \bar y\|_Y^2 - \frac{1}{2\gamma_g}\|(K^{*\approx} - K^*)y^k\|_Z^2 \\
&\geq \gamma\|x^{k+1} - \bar x\|_M^2 - \varepsilon_k(\bar x).
\end{aligned}
$$

Therefore, (4.6) holds with the stated choices. Moreover, we have $\sum_{k=0}^{N-1} p^{k-N} \leq 1/(p-1)$ for any $p \in (1, 1+2\gamma]$, verifying (4.8) and consequently Assumption 4.11 (a). □

Remark 4.18 (Stochastic optimisation methods). These can be approached through lifting: we take X as a space of random variables on a space $\mathscr{X}$, set $F(x) = \mathbb{E}[\mathscr{F} \circ x]$ for a function $\mathscr{F}$ on $\mathscr{X}$, $Mx := \mathbb{E}^*\mathscr{M}\mathbb{E}[x]$ for $\mathscr{M} \in \mathbb{L}(\mathscr{X}; \mathscr{X}^*)$, etc., where $\mathbb{E}$ is the expectation. Without inexactness, this produces parallel copies of the very same optimisation method for every random event ω. We then model the random choices made on every step through inexactness, subject to Assumptions 4.8, 4.10 and 4.11.

4.3 CONVERGENCE OF SUBDIFFERENTIALS AND QUASI-MONOTONICITY OF VALUES

We first show the potentially global convergence of subdifferentials; see Remark 4.9. When $\Xi = 0$, this could be followed by the Kurdyka–Łojasiewicz property to show function value convergence, and, afterwards, either by a growth condition or, in finite dimensions, a finite-length argument based on (4.20) and [1, proof of Lemma 2.6] to show iterate convergence. As the property can easily be verified only in finite dimensions (for semi-algebraic functions), we prefer a more direct approach.

Theorem 4.19. *If Assumption 4.8 holds, then* $x^k \in \Omega$ *and*

$$\mathcal{G}(x^{k+1}; x^k) + \eta\|x^{k+1} - x^k\|_M^2 \le \varepsilon_{\text{desc},k} \quad \text{for all} \quad k \in \mathbb{N}. \tag{4.20}$$

We also have $\inf_{x^* \in H(x^{k+1})} \|x^*\|_{X^*} \to 0$ *if, moreover, Assumption 4.10 holds and*

$$\inf_{N \in \mathbb{N}} \sum_{k=0}^{N-1} \left(\mathcal{G}(x^{k+1}; x^k) + \tilde{\eta}\|x^{k+1} - x^k\|_M^2 \right) > -\infty \quad \text{for some } \tilde{\eta} < \eta. \tag{4.21}$$

Proof. By the implicit algorithm (4.3), the properties of Fenchel conjugates (e.g., [6, Lemma 5.7]) and $-M(x^{k+1} - x^k) =: \tilde{\partial}_{k+1} \in \partial \left(\frac{1}{2}\|\,\cdot\,\|_M^2\right)(x^{k+1} - x^k)$, we have

$$(\|\,\cdot\,\|_M^2)^*(2\tilde{\partial}_{k+1}) = 2\left(\frac{1}{2}\|\,\cdot\,\|_M^2\right)^*(\tilde{\partial}_{k+1}) = \|x^{k+1} - x^k\|_M^2 = -\langle \tilde{\partial}_{k+1} | x^{k+1} - x^k \rangle_{X^*,X}. \tag{4.22}$$

If $\{x^j\}_{j=0}^{N-1} \subset \Omega$, Assumption 4.8 (i) thus yields for all $k = 0, \ldots, N-1$ that

$$\begin{aligned} \mathcal{G}(x^{k+1}; x^k) &= \mathcal{G}(x^{k+1}; x^k) - \langle \tilde{\partial}_{k+1} | x^{k+1} - x^k \rangle_{X^*,X} - \|x^{k+1} - x^k\|_M^2 \\ &\le \varepsilon_{\text{desc},k} - \frac{1}{2}\|x^{k+1} - x^k\|_{2M-\check{\Lambda}}^2 \le \varepsilon_{\text{desc},k} - \eta\|x^{k+1} - x^k\|_M^2. \end{aligned} \tag{4.23}$$

Summing over all such k, and using Assumption 4.8 (ii), it follows

$$\sum_{k=0}^{N-1} \mathcal{G}(x^{k+1}; x^k) + \sum_{k=0}^{N-1} \eta(\|\,\cdot\,\|_M^2)^*(2\tilde{\partial}_{k+1}) = \sum_{k=0}^{N-1} \mathcal{G}(x^{k+1}; x^k) + \sum_{k=0}^{N-1} \eta\|x^{k+1} - x^k\|_M^2 \le r_{\text{desc}}.$$

From Assumption 4.8 (iii), it now follows that $x^N \in \Omega$. Since, by the same assumption, $x^0 \in \Omega$, induction establishes (4.20) and $x^k \in \Omega$ for all $k \in \mathbb{N}$. Using (4.21), we, moreover, deduce $\sup_{N \in \mathbb{N}} \sum_{k=0}^{N-1} \|x^{k+1} - x^k\|_M^2 < \infty$ and $(\|\,\cdot\,\|_M^2)^*(2\tilde{\partial}_{k+1}) \to 0$. Let $c \ge \|M\|_{\mathbb{L}(X;X^*)}$. By $\|\,\cdot\,\|_M^2 \le c\|\,\cdot\,\|_X^2$ and the properties of conjugates (e.g., [6, Lemmas 5.4 and 5.7]),

$$\frac{4}{c}\|\tilde{\partial}_{k+1}\|_{X^*}^2 = c\|2\tilde{\partial}_{k+1}/c\|_{X^*}^2 = (c\|\,\cdot\,\|_X^2)^*(2\tilde{\partial}_{k+1}) \le (\|\,\cdot\,\|_M^2)^*(2\tilde{\partial}_{k+1}).$$

Thus also $\|\tilde{\partial}_{k+1}\|_{X^*} \to 0$. Assumption 4.10 proves that $\inf_{x^* \in H(x^{k+1})} \|\tilde{\partial}_{k+1} - x^*\|_{X^*} \to 0$. Hence an application of the triangle inequality establishes $\inf_{x^* \in H(x^{k+1})} \|x^*\|_{X^*} \to 0$. ◻

Remark 4.20 (Bregman divergences). The argument of Theorem 4.19 extends to algorithms where the M-seminorm for $M \in \mathbb{L}(X; X^*)$ is replaced by a Bregman divergence B_M generated by some convex $M : X \to \overline{\mathbb{R}}$. Recalling the definition

$$B_M^\omega(x, z) := M(z) - M(x) - \langle \omega | z - x \rangle \quad (\omega \in \partial M(x); x, z \in X), \tag{4.24}$$

in the algorithm (4.3), we would impose $\tilde{\partial}_{k+1} = \omega^{k+1} - \omega^k \in \partial_2 B_M^{\omega^k}(x^k, x^{k+1})$ for a given $\omega^k \in \partial M(x^k)$ and some $\omega^{k+1} \in \partial M(x^{k+1})$. Using the Fenchel–Young identity, we could then replace (4.22) by

$$\begin{aligned} B_{M^*}^{x^{k+1}}(\omega^{k+1}, \omega^k) = B_M^{\omega^k}(x^k, x^{k+1}) &= M(x^{k+1}) - M(x^k) - \langle \omega^k | x^{k+1} - x^k \rangle \\ &\le \langle \omega^{k+1} - \omega^k | x^{k+1} - x^k \rangle = -\langle \tilde{\partial}_{k+1} | x^{k+1} - x^k \rangle_{X^*,X}. \end{aligned}$$

Thus repeating the arguments of the theorem would establish both $B_M^{\omega^k}(x^k, x^{k+1}) \to 0$ as well as $B_{M^*}^{x^{k+1}}(\omega^{k+1}, \omega^k) = B_{M^*}^{x^{k+1}}(\omega^k + \tilde{\partial}_{k+1}, \omega^k) \to 0$. A variant of Assumption 4.10 could then establish a form of convergence for $H(x^{k+1})$.

Example 4.21 (Forward-backward splitting subdifferential convergence). For the (inexact) forward-backward splitting of Example 4.1, the condition (4.21) amounts to $\inf[F+G] > -\infty$. Subject to Assumptions 4.8 and 4.10 (see Examples 4.13 and 4.15), Theorem 4.19 establishes the monotonicity of function values, as well as the convergence of subdifferentials to zero, $\inf_{x^*\in\partial G(x^{k+1})} \|F'(x^{k+1}) + x^*\| \to 0$.

Example 4.22 (PDPS "co-convergence"). Because Ξ is not cyclically monotone (see [22, Chapter 24]), we see no way in general[2] for the PDPS to satisfy (4.21). However, we can monitor potential convergence failure by setting an expected lower bound on

$$\sum_{k=0}^{N-1} \mathcal{G}(x^{k+1}; x^k) = [F+G](x^N) - [F+G](x^0) - \sum_{k=0}^{N-1} \langle \Xi x^{k+1} | x^k \rangle.$$

In fact, if $\inf F + G > -\infty$, we only need to ensure that the latter sum stay above a chosen bound, without having to calculate potentially costly function values.

4.4 non-escape, quasi-féjer monotonicity, linear convergence

The next lemma is essential for all our strong convergence results. The proof is standard; see, e.g., [6, Chapter 15] for the case $\varepsilon_k(\bar x) = 0$ and $\Xi = 0$. Observe that (4.25) with the triangle inequality may be used to again prove Assumption 2.2 (i) for multilevel methods.

Lemma 4.23. *Suppose Assumption 4.11 holds at $\bar x \in X$. Then $x^k \in \mathbb{O}_M(\bar x, \delta) \subset \Omega_{\bar x}$ for all $k \in \mathbb{N}$, and the sequence is (p-strongly) quasi-Féjer, i.e.,*

$$\frac{p}{2}\|x^{k+1} - \bar x\|_M^2 \le \frac{1}{2}\|x^k - \bar x\|_M^2 + \varepsilon_k(\bar x). \tag{4.25}$$

Moreover, $\sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} p^{k-N} \|x^{k+1} - x^k\|_M^2 < \infty$ if $\eta > 0$.

Proof. We first treat Assumption 4.11 option (a). Fix $N \in \mathbb{N}$ and suppose $\{x^j\}_{j=0}^{N-1} \subset \Omega_{\bar x}$. Observe that $\langle \Xi x | x \rangle = 0$ for all $x \in X$ by the skew-adjointness of Ξ. Since $0 \in H(\bar x)$, using (4.6) in the implicit algorithm (4.3), we thus get

$$-\langle M(x^{k+1} - x^k) | x^{k+1} - \bar x \rangle_{X^*,X} \ge \gamma \|x^{k+1} - \bar x\|_M^2 - \frac{1}{2}\|x^{k+1} - x^k\|_{\check\Lambda}^2 - \varepsilon_k(\bar x)$$

for all $k \in \{0, \ldots, N-1\}$. By $\check\Lambda \le (1-\eta)M$ and the Pythagoras' identity (see [30, (2.3)])

$$\langle M(x - z) | x - \bar x \rangle = \frac{1}{2}\|x - z\|_M^2 + \frac{1}{2}\|x - \bar x\|_M^2 - \frac{1}{2}\|z - \bar x\|_M^2 \quad (x, z, \bar x \in X),$$

we obtain

$$\frac{1}{2}\|x^k - \bar x\|_M^2 \ge \frac{\eta}{2}\|x^{k+1} - x^k\|_M^2 + \frac{1+2\gamma}{2}\|x^{k+1} - \bar x\|_M^2 - \varepsilon_k(\bar x). \tag{4.26}$$

[2] If G is smooth and F is appreciably strongly convex, then, for an exact method, we can for some $q^{k+1} \in \partial G(x^{k+1})$ use (4.23) to expand and estimate $\mathcal{G}(x^{k+1}; x^k) + \tilde\eta \|x^{k+1} - x^k\|_M^2 = [F+G](x^{k+1}) - [F+G](x^k) - \langle F'(x^k) + q^{k+1}, x^{k+1} - x^k \rangle - (1 - \tilde\eta)\|x^{k+1} - x^k\|_M^2 \ge 0$. This result is, unfortunately, unhelpful.

Using $1+2\gamma \ge p$, multiplying by p^k, and summing over $k=0,\dots,N-1$ yields

$$\frac{1}{2}\|x^0-\bar x\|_M^2 + \sum_{k=0}^{N-1} p^k \varepsilon_k(\bar x) \ge \sum_{k=0}^{N-1} \frac{\eta p^k}{2}\|x^{k+1}-x^k\|_M^2 + \frac{p^N}{2}\|x^N-\bar x\|_M^2. \tag{4.27}$$

Multiplying by $p^{-N} \le 1$ and using $x^0 \in \mathbb{O}_M(\bar x, \sqrt{\delta^2-2r_p})$ and (4.8), it follows

$$\frac{\delta^2}{2} = \frac{\delta^2-2r_1}{2} + r_1 > \sum_{k=0}^{N-1} \frac{\eta p^{k-N}}{2}\|x^{k+1}-x^k\|_M^2 + \frac{1}{2}\|x^N-\bar x\|_M^2. \tag{4.28}$$

Hence $x^N \in \mathbb{O}_M(\bar x,\delta)$ if $p \ge 1$, while the alternative $\delta=\infty$ is obvious. Since $x^0 \in \Omega_{\bar x}$ by Assumption 4.11, an inductive argument shows that $x^k \in \mathbb{O}_M(\bar x,\delta) \subset \Omega_{\bar x}$ for all $k\in\mathbb{N}$, justifying the above steps. Finally, (4.26) shows (4.25), while $\sup_{N\in\mathbb{N}} \sum_{k=0}^{N-1} p^{k-N}\|x^{k+1}-x^k\|_M^2 < \infty$ follows from (4.28) and $\eta>0$.

Regarding option Assumption 4.11 (b), arguing as above with (4.7) in place of (4.6), we get in place of (4.26) the estimate

$$\frac{1}{2}\|x^k-\bar x\|_M^2 \ge \mathcal{G}(x^{k+1};\bar x) + \frac{\eta}{2}\|x^{k+1}-x^k\|_M^2 + \frac{1+\gamma}{2}\|x^{k+1}-\bar x\|_M^2 - \varepsilon_k(\bar x). \tag{4.29}$$

Using $\inf_{x\in\mathbb{O}_M(\delta,\bar x)} \mathcal{G}(x;\bar x) \ge 0$, we proceed (with $1+\gamma \ge p$) as in option (a) to establish (4.28), and from there onwards. ◻

A closer look at (4.27) immediately yields linear convergence if $p>1$.

Corollary 4.24. *Suppose Assumption 4.11 holds at $\bar x \in X$ with $p>1$. Then $\|x^N-\bar x\|_M^2 \to 0$ at the rate $O(p^{-N})$.*

4.5 LOCAL CONVERGENCE OF FUNCTION VALUES

We now proceed to function values and duality gaps. The idea of possibly assuming both Assumption 4.11 (a) and a relaxed version of (b), as an alternative to just the latter, is to be able to study descent at non-minimising critical points. For simplicity, we only treat sublinear convergence.

Theorem 4.25. *Suppose Assumption 4.11 holds at $\bar x \in X$ and, for a non-empty set $\hat X \subset X$, (4.7) holds for all $\hat x \in \hat X$ with $\breve\Lambda = \breve\Lambda_{\hat x} \le M$, $\gamma=\gamma_{\hat x} \ge 0$, and $\Omega_{\hat x} \supset \mathbb{O}_M(\bar x,\delta)$. Then*

$$\sup_{\hat x\in\hat X} \sum_{k=0}^{N-1} \mathcal{G}(x^{k+1};\hat x) \le \sup_{\hat x\in\hat X}\left(\frac{1}{2}\|x^0-\hat x\|_M^2 + \sum_{k=0}^{N-1}\varepsilon_{\mathrm{desc},k}(\hat x)\right) \quad \text{for all} \quad N\in\mathbb{N}. \tag{4.30}$$

If $\Xi=0$ and Assumption 4.8 holds[3]*, then, for all $N\in\mathbb{N}$,*

$$[F+G](x^N) \le \inf_{\hat x\in\hat X}[F+G](\hat x) + \sup_{\hat x\in\hat X}\left(\frac{1}{2N}\|x^0-\hat x\|_M^2 + \sum_{k=0}^{N-1}\left(\frac{1}{N}\varepsilon_{\mathrm{desc},k}(\hat x) + \frac{k+1}{N}\varepsilon_{\mathrm{desc},k}\right)\right). \tag{4.31}$$

Proof. Lemma 4.23 shows for all $k\in\mathbb{N}$ that $x^k \in \mathbb{O}_M(\bar x,\delta) \subset \bigcap_{\hat x\in\hat X}\Omega_{\hat x}$. Hence, for any $\hat x\in\hat X$, we may follow the proof of the lemma for case (b) of Assumption 4.11 to establish (4.29) for $\bar x = \hat x$. To reach this point, the assumption $\inf_{x\in\mathbb{O}_M(\delta,\bar x)}\mathcal{G}(x;\bar x) \ge 0$ was not yet needed. Now, summing (4.29) over $k=0,\dots,N-1$, we obtain

$$\frac{1}{2}\|x^0-\hat x\|_M^2 + \sum_{k=0}^{N-1}\varepsilon_{\mathrm{desc},k}(\hat x) \ge \sum_{k=0}^{N-1}\mathcal{G}(x^{k+1};\hat x) + \frac{1}{2}\|x^N-\hat x\|_M^2. \tag{4.32}$$

[3] Since the proof of the present Theorem 4.25 shows that $x^k \in \mathbb{O}_M(\hat x,\delta)$ for all $k\in\mathbb{N}$, to prove the required (4.20), it would be enough to assume that just Assumption 4.8 (i) holds with $\Omega \supset \mathbb{O}_M(\hat x,\delta)$.

Taking the supremum over $\hat{x} \in \hat{X}$, this establishes (4.30).

Suppose then that $\Xi = 0$ and Assumption 4.8 holds. Theorem 4.19 now establishes (4.20), i.e., the quasi-monotonicity $[F+G](x^{k+1}) \le [F+G](x^k) + \varepsilon_{\text{desc},k}$. Repeatedly using this and $\mathcal{G}(x^{k+1};\hat{x}) = [F+G](x^{k+1}) - [F+G](\hat{x})$ in (4.32), and dividing by N, we obtain (4.31). □

We next specialise the result to the PDPS of Example 4.2. Besides inexactness, as a novelty compared to [4, 5, 16, 10], subject to h^* having a bounded domain, we get an estimate on the convex envelope of the objective, i.e., the Fenchel biconjugate. In non-reflexive spaces, we define the latter as a function in X instead of X^{**} by taking first the conjugate and then the equivalently defined preconjugate: $h^{**} := (h^*)_*$.

Corollary 4.26. *Assume the setup of Example 4.2 and Assumption 4.3 for some $\tau, \sigma, \lambda > 0$, as well as that Theorem 4.16* (ii) and (iii) *hold for $p = 1$ at some $\bar{z} \in Z$ with $\mathbb{O}_{M_z}(\bar{z}, \delta_z) \subset \Omega_{\bar{z}}$ for some $\delta_z > 0$. Also suppose that that* dom h^* *is bounded, $0 \in H(\bar{x})$ for some $\bar{x} \in \{\bar{z}\} \times \operatorname{dom} h^*$, and that the condition (4.16) on the initial iterate holds. Then, for the* ergodic iterates *$\tilde{z}^N := \frac{1}{N}\sum_{k=0}^{N-1} z^k$, for all $N \in \mathbb{N}$, we have*

$$[f + g + h \circ K]^{**}(\tilde{z}^N) \le [f + g + h \circ K](\bar{z}) + \sup_{\hat{y} \in \operatorname{dom} h^*} \frac{1}{2N}\|(z^0, y^0) - (\bar{z}, \hat{y})\|_M^2 + \frac{\sum_{k=0}^{N-1} e_{1,k}(z^{k+1})}{2\tilde{\gamma}\lambda N}.$$

Here $[f + g + h \circ K](\bar{z}) = [f + g + h \circ K]^{**}(\bar{z})$ if $\bar{z}$ is a global minimiser of $f + g + h \circ K$.

Proof. Theorem 4.16 (ii) proves Assumption 4.11 option (a) at $\bar{x}$. Likewise, Theorem 4.16 (iii) shows (4.7) and $\Omega_{\hat{x}} := \Omega_{\bar{z}} \times \operatorname{dom} h^* \supset \mathbb{O}_M(\bar{x}, \delta)$ at any $\hat{x} \in \hat{X} := \{\bar{z}\} \times \operatorname{dom} h^*$. Theorem 4.25 now establishes (4.30), whose left-hand-side we still have to estimate.

With the expression of Example 4.7 for the gap, we expand and estimate using the definition of the Fenchel (bi)conjugate and $h^{**} = h$ as well as $[f+g]^{**} \le f + g$ that

$$\begin{aligned}\mathcal{G}(x^{k+1};\bar{x}) &= ([f+g](z^{k+1}) + \langle Kz^{k+1}|\hat{y}\rangle - h^*(\hat{y})) - ([f+g](\bar{z}) + \langle K\bar{z}|y^{k+1}\rangle - h^*(y^{k+1})) \\ &\ge \left([f+g]^{**}(z^{k+1}) + \langle Kz^{k+1}|\hat{y}\rangle - h^*(\hat{y})\right) - N[f+g+h\circ K](\bar{z}).\end{aligned}$$

Summing over $k \in \{0, \ldots, N-1\}$, taking the supremum over $\hat{y} \in \operatorname{dom} h^*$, and using Jensen's inequality, therefore

$$\sup_{\hat{y} \in \operatorname{dom} h^*} \sum_{k=0}^{N-1} \mathcal{G}(x^{k+1};\bar{x}) \ge N[(f+g)^{**} + h \circ K](\tilde{z}^N) - N[f+g+h\circ K](\bar{z}).$$

Denoting the infimal convolution by $\square$, we have

$$f + g + h \circ K \ge [f + g + h \circ K]^{**} = ((f+g)^* \,\square\, [h \circ K]^*)^* = (f+g)^{**} + h \circ K.$$

Moreover, the inequality is an equality at a global minimiser (or if f is convex). Now the claim follows from (4.30). □

Remark 4.27. Taking $p > 1$ in the proof of Corollary 4.26, linear convergence rates could be obtained as in Corollary 4.24 for the iterates.

4.6 Weak convergence

We next prove weak-∗ convergence of the iterates. For this, we call the self-adjoint and positive semi-definite preconditioner $M \in \mathbb{L}(X; X^*)$ *weak-∗ admissible* if $\|x^k\|_M \to 0$ implies $Mx^k \to 0$.

Example 4.28. Suppose $M = A^*A$ for some $A \in \mathbb{L}(X; V)$ for a Hilbert space V. Then the M-semi-norm convergence clearly implies $Ax^k \to 0$, and consequently $Mx^k \to 0$. Thus M is weak-∗ admissibility. In Hilbert spaces every positive-definite self-adjoint operator has such a square root A with $W = X$. For a convolution-based construction in the space of Radon measures, see [30, Theorem 2.4].

Theorem 4.29. *Suppose Assumptions 4.10 and 4.11 hold with $p = 1$ and $\eta > 0$ at some $\bar{x} = \hat{x} \in H^{-1}(0)$, and that either Assumption 4.11 (a) or (b) (only the item, not the entire assumption) holds with $\mathbb{O}_M(\bar{x}, \delta) \subset \Omega_{\hat{x}}$ and $\sum_{k=0}^{\infty} \varepsilon_k(\hat{x}) < \infty$ at all $\hat{x} \in \hat{X} := H^{-1}(0) \cap \mathbb{O}_M(\bar{x}, \delta)$. Also suppose that X is the dual space of a separable normed space X_*, the preconditioner M is weak-∗ admissible, and F is either convex or F' is weak-∗-to-strong continuous. Then $x^k \overset{*}{\rightharpoonup} \hat{x}$ weakly-∗ for some $\hat{x} \in \hat{X}$.*

Proof. Lemma 4.23 proves that $x^k \in \mathbb{O}_M(\bar{x}, \delta)$ for all $k \in \mathbb{N}$, as well as that $\sup_{N \in N} \sum_{k=0}^{N-1} \|x^{k+1} - x^k\|_M^2 < \infty$. The latter establishes $\|x^{k+1} - x^k\|_M \to 0$, and through weak-∗ admissibility and (4.3) that $\tilde{\partial}_{k+1} = -M(x^{k+1} - x^k) \to 0$ strongly in X^*. Moreover, Assumption 4.10 yields $\|\tilde{\partial}_{k+1} - x_{k+1}^*\|_{X^*} \to 0$ for some $x_{k+1}^* \in H(x^{k+1})$. Consequently $x_{k+1}^* \to 0$. Since $x^k \in \mathbb{O}_M(\bar{x}, \delta) \subset \Omega_{\hat{x}}$, as in the proof of Lemma 4.23, we show the quasi-Féjer monotonicity (4.25) for all $\hat{x} \in \hat{X}$ and $k \in \mathbb{N}$.

Suppose then that $x^{k_j+1} \overset{*}{\rightharpoonup} \hat{x}$ for a subsequence $\{k_j\}_{j \in \mathbb{N}} \subset \mathbb{N}$ and a $\hat{x} \in X$. We want to show that $\hat{x} \in \hat{X}$. We consider two cases:

1. If F is convex, H is maximally monotone[4], hence weak-∗-to-strong outer semicontinuous. Now $x^{k_j+1} \overset{*}{\rightharpoonup} \hat{x}$ and $H(x^{k_j+1}) \ni x_{k_j+1}^* \to 0$ obliges $0 \in H(\hat{x})$.

2. Suppose then that F' is weak-∗-to-strong continuous. Now still $P : x \mapsto \partial G(x) + \Xi x$ is maximally monotone[4], hence weak-∗-to-strong outer semicontinuous. We have $P(x^{k_j+1}) \ni x_{k_j+1}^* - F'(x^{k_j+1}) \to -F'(\hat{x})$ strongly in X^*, as well as $x^{k_j+1} \overset{*}{\rightharpoonup} \hat{x}$, so we must have $-F'(\hat{x}) \in P(\hat{x})$. But this again says $0 \in H(\hat{x})$.

Thus every weak-∗ limiting point $\hat{x}$ of $\{x^k\}_{k \in \mathbb{N}}$ satisfies $0 \in H(\hat{x})$. But, since $x^k \in \mathbb{O}_M(\bar{x}, \delta)$ for all $k \in \mathbb{N}$, also $\hat{x} \in \mathbb{O}_M(\bar{x}, \delta)$. This proves that $\hat{x} \in \hat{X}$. Since, by assumption, $\sum_{k=0}^{\infty} \varepsilon_k(\hat{x}) < \infty$ for all $\hat{x} \in \hat{X}$, the quasi-Féjer monotonicity (4.25) with the quasi-Opial's Lemma A.2 finishes the proof. □

Example 4.30. In the setting of Section 2 and Theorem 2.14, the weak-∗-to-strong continuity of F' can be achieved, for example, when $F(x) = \frac{1}{2}\|S(x) - b\|^2$ for a Lipschitz and bounded S with finite-dimensional range.

APPENDIX A OPIAL'S LEMMA FOR QUASI-FÉJER MONOTONICITY

Here we prove a generalisation of Opial's lemma [20] for quasi-Féjer monotonicity, i.e, Féjer monotonicity with an additive error term. We prove it in normed spaces for Bregman divergences (4.24), as they add no extra difficulties. In an even more general variable-metric framework, a similar result is also proved in [31, Proposition 2.7]. Our simplified proof follows the outline of that in [6], and is nearly identical to the one in [30], where the errors took a more specific form.

For the proof, we recall the following deterministic version of the results of [21]:

[4] That the additive skew-adjoint term Ξ does not destroy maximal monotonicity, can be proved completely analogously to the Hilbert space case in [6, Lemma 9.9].

Lemma A.1. *Let* $\{a_k\}_{k\in\mathbb{N}}$, $\{b_k\}_{k\in\mathbb{N}}$, $\{c_k\}_{k\in\mathbb{N}}$, *and* $\{d_k\}_{k\in\mathbb{N}}$ *be non-negative and* $a_{k+1} \le a_k(1+b_k)+c_k-d_k$ *for all* $k \in \mathbb{N}$. *If* $\sum_{k=0}^\infty b_k < \infty$ *and* $\sum_{k=0}^\infty c_k < \infty$, *then (i)* $\lim_{k\to\infty} a_k$ *exists and is finite; and (ii)* $\sum_{k=0}^\infty d_k < \infty$.

Lemma A.2. *Let either* X *be the dual space of a corresponding separable normed space* X_*, *or, alternatively, let* X *be reflexive. Also let* $M : X \to \mathbb{R}$ *be convex, proper, and Gâteaux differentiable with* $M' : X \to X_*$ *weak-∗-to-weak continuous. Finally, let* $\hat{X} \subset X$ *be non-empty and* $\{e_k\}_{k\in\mathbb{N}} \in \mathbb{R}$. *If*

(i) all weak-∗ limit points of $\{x^k\}_{k\in\mathbb{N}}$ *belong* $\hat{X}$*;*

(ii) $B_M(x^{k+1}, \bar{x}) \le B_M(x^k, \bar{x}) + e_k(\bar{x})$ *for some* $e_k(\bar{x}) \ge 0$ *for all* $\bar{x} \in \hat{X}$ *and* $k \in \mathbb{N}$*; and*

(iii) $\sum_{k=0}^\infty e_k(\bar{x}) < \infty$ *for all* $\bar{x} \in \hat{X}$*;*

then all weak-∗ limit points of $\{x^k\}_{k\in\mathbb{N}}$ *satisfy* $\hat{x}, \bar{x} \in \hat{X}$ *and*

$$\langle M'(\hat{x}) - M'(\bar{x}) | \hat{x} - \bar{x} \rangle = 0. \tag{A.1}$$

If $\{x^k\}_{k\in\mathbb{N}} \subset X$ *is bounded, then such a limit point exists. If, in addition to all the previous assumptions, (A.1) implies* $\hat{x} = \bar{x}$ *(such as when* M *is strongly monotone), then* $x^k \overset{*}{\rightharpoonup} \hat{x}$ *weakly-∗ in* X *for some* $\hat{x} \in \hat{X}$.

Proof. Let $\bar{x}$ and $\hat{x}$ be weak-∗ limit points of $\{x^k\}_{k\in\mathbb{N}}$. Since Bregman divergences $B_M \ge 0$ for convex M, the conditions (ii) and (iii) establish the assumptions of Lemma A.1 for $a_k = B_M(x^k; \bar{x})$, $b_k = 0$, $c_k = e_k(\bar{x})$, and $d_k = 0$. It follows that $\{B_M(x^k; \bar{x})\}_{k\in\mathbb{N}}$ is convergent. Likewise we establish that $\{B_M(x^k; \hat{x})\}_{k\in\mathbb{N}}$ is convergent. Therefore, by the obvious three-point identity for Bregman divergences (see, e.g., [28]),

$$\langle M'(x^k) - M'(\hat{x}) | \bar{x} - \hat{x} \rangle = B_M(x^k; \hat{x}) - B_M(x^k; \bar{x}) + B_M(\hat{x}; \bar{x}) \to c \in \mathbb{R}.$$

Since $\bar{x}$ and $\hat{x}$ are a weak-∗ limit point, there exist subsequences $\{x^{k_n}\}_{n\in\mathbb{N}}$ and $\{x^{k_m}\}_{m\in\mathbb{N}}$ with $x^{k_n} \rightharpoonup \bar{x}$ and $x^{k_m} \rightharpoonup \hat{x}$. By the weak-∗-to-weak continuity of $M' : X \to X_*$, (A.1) follows from

$$\langle M'(\bar{x}) - M'(\hat{x}) | \bar{x} - \hat{x} \rangle = \lim_{n\to\infty} \langle M'(x^{k_n}) - M'(\hat{x}) | \bar{x} - \hat{x} \rangle = c = \lim_{m\to\infty} \langle M'(x^{k_m}) - M'(\hat{x}) | \bar{x} - \hat{x} \rangle = 0.$$

If $\{x^k\}_{k\in\mathbb{N}}$ is bounded, and X is the dual space of some separable normed space X_*, it contains a weakly-∗ convergent subsequence by the Banach–Alaoglu theorem, so a limit point exists as claimed. If X is reflexive, the Eberlein–Šmulyan theorem establishes the same result. Hence, if (A.1) implies $\bar{x} = \hat{x}$, then every convergent subsequence of $\{x^k\}_{k\in\mathbb{N}}$ has the same weak limit. It lies in $\hat{X}$ by (i). The final claim now follows from a standard subsequence–subsequence argument: Assume to the contrary that there exists a subsequence of $\{x^k\}_{k\in\mathbb{N}}$ not convergent to $\hat{x}$. Then the above argument provides a further subsequence converging to $\hat{x}$. This contradicts the fact that any subsequence of a convergent sequence converges to the same limit. □

REFERENCES

[1] H. Attouch, J. Bolte, and B. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods, *Mathematical Programming* 137 (2013), 91–129, doi:10.1007/s10107-011-0484-9.

[2] R. J. Baraldi and D. P. Kouri, A proximal trust-region method for nonsmooth optimization with inexact function and gradient evaluations, *Mathematical Programming* 201 (2022), 55––598, doi:10.1007/s10107-022-01915-3.

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

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

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

[6] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, 2020, arXiv:2001.00216. Submitted.

[7] J. C. De Los Reyes, *Numerical PDE-Constrained Optimization*, SpringerBriefs in Optimization, Springer International Publishing, 2015, doi:10.1007/978-3-319-13395-9.

[8] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, *Mathematical Programming* 146 (2013), 37–75, doi:10.1007/s10107-013-0677-5.

[9] P. E. Dvurechensky, A gradient method with inexact oracle for composite nonconvex optimization, *Computer Research and Modeling* 14 (2022), 321–334, doi:10.20537/2076-7633-2022-14-2-321-334.

[10] Y. Gao and W. Zhang, An alternative extrapolation scheme of PDHGM for saddle point problem with nonlinear function, *Computational Optimization and Applications* 85 (2023), 263–291, doi:10.1007/s10589-023-00453-8.

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic press, 2014.

[12] 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.

[13] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, number 23 in Mathematical Modelling: Theory and Applications, Springer Netherlands, 2009, doi:10.1007/978-1-4020-8839-1.

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

[15] D. A. Lorenz and F. Schneppe, Chambolle–Pock's Primal-Dual Method with Mismatched Adjoint, *Applied Mathematics and Optimization* 87 (2023), doi:10.1007/s00245-022-09933-5.

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

[17] T. Möllenhoff, E. Strekalovskiy, M. Moeller, and D. Cremers, The primal-dual hybrid gradient method for semiconvex splittings, *SIAM Journal on Imaging Sciences* 8 (2015), 827–857, doi:10.1137/140976601.

[18] Y. Nabou, F. Glineur, and I. Necoara, Proximal gradient methods with inexact oracle of degree q for composite optimization, *Optimization Letters* (2024), doi:10.1007/s11590-024-02118-9.

[19] P. Ochs, Unifying Abstract Inexact Convergence Theorems and Block Coordinate Variable Metric iPiano, *SIAM Journal on Optimization* 29 (2019), 541–570, doi:10.1137/17m1124085.

[20] Z. Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bulleting of the American Matheatical Society* 73 (1967), 591–597, doi:10.1090/s0002-9904-1967-11761-0.

[21] H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, *Optimizing Methods in Statistics* (1971), 233–257, doi:10.1016/b978-0-12-604550-5.50015-8.

[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1972.

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

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

[25] E. Suonperä and T. Valkonen, General single-loop methods for bilevel parameter learning, 2024, arXiv:2408.08123. submitted.

[26] 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.

[27] 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.

[28] 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.

[29] T. Valkonen, Predictive online optimisation with applications to optical flow, *Journal of Mathematical Imaging and Vision* 63 (2021), 329–355, doi:10.1007/s10851-020-01000-4, arXiv:2002.03053.

[30] T. Valkonen, Proximal methods for point source localisation, *Journal of Nonsmooth Analysis and Optimization* 4 (2023), 10433, doi:10.46298/jnsao-2023-10433, arXiv:2212.02991.

[31] Q. Van Nguyen, Variable quasi-Bregman monotone sequences, *Numerical Algorithms* 73 (2016), 1107–1130, doi:10.1007/s11075-016-0132-9.

[32] M. Yan and Y. Li, On the Improved Conditions for Some Primal-Dual Algorithms, *Journal of Scientific Computing* 99 (2024), doi:10.1007/s10915-024-02537-x.